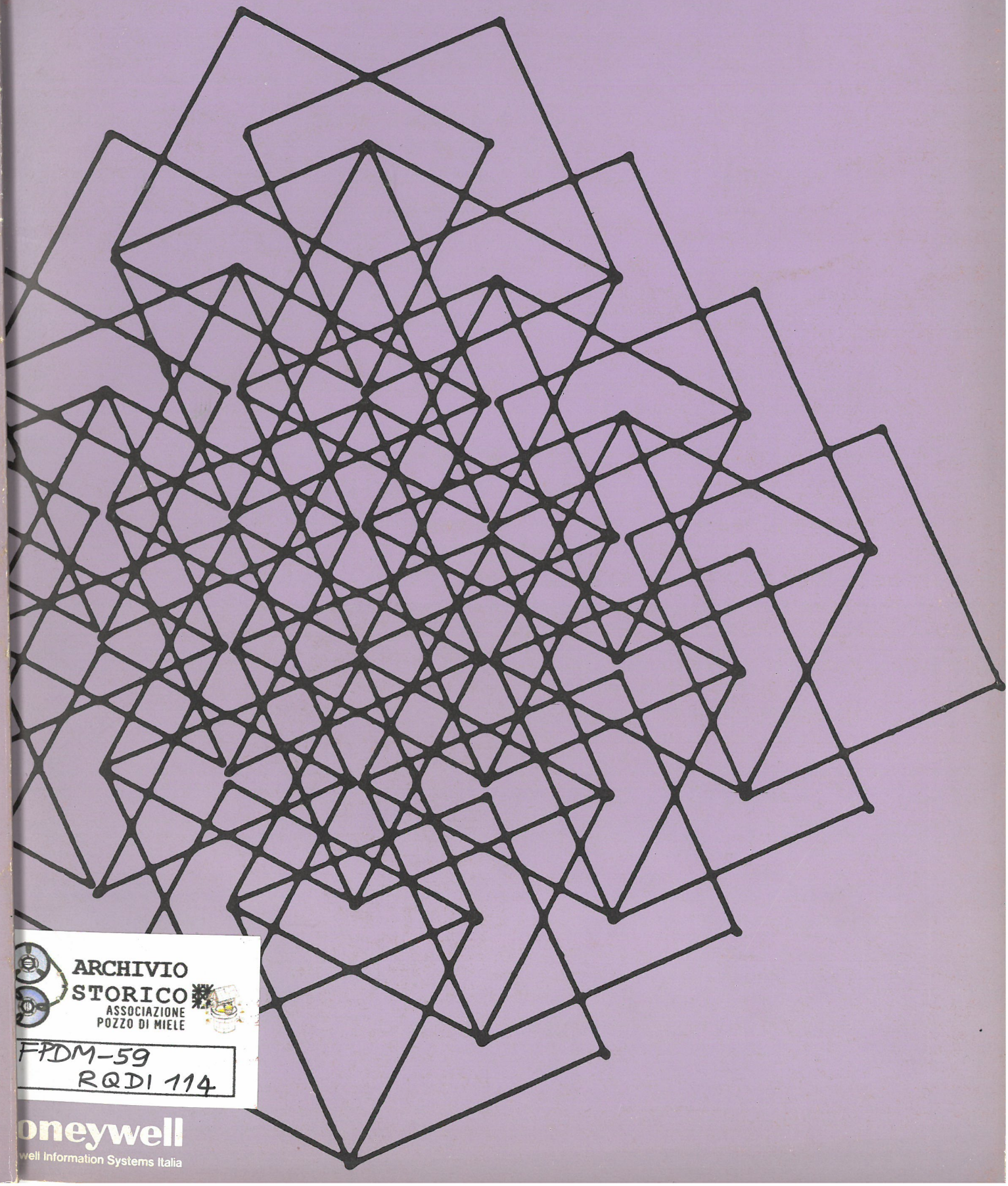
Quaderni di informatica 16

Rassegna di tecnologia ed applicazioni degli elaboratori elettronici - Anno VIII - Numero 2 - 1981



ARCHIVIO STORICO
ASSOCIAZIONE POZZO DI MIELE
FPDM-59
RQDI 114

oneywell
well Information Systems Italia

Honeywell
Honeywell Information Systems Italia

Quaderni di informatica

Rassegna di tecnologia ed applicazioni degli elaboratori elettronici
Anno VIII - Numero 2 - 1981

Sommario

ADA: un linguaggio per gli anni '80

Vincitore di una lunga e articolata selezione a livello mondiale, ADA costituisce un fondamentale approccio ai linguaggi di programmazione. Come anche il nome vuole ricordare, questo linguaggio ha una origine europea; esso è stato infatti sviluppato dalla consociata francese della Honeywell.
G.L. CASTELLI

Supercalcolatori per applicazioni speciali

La velocità di calcolo dei più potenti elaboratori attuali può risultare insufficiente in determinate applicazioni. L'articolo esamina nuove soluzioni strutturali potenzialmente atte a realizzare le velocità richieste.
J.W. BOWRA

Connessioni con fibre ottiche nei sistemi di elaborazione

L'impiego di fibre ottiche per la trasmissione di segnali presenta sostanziali vantaggi rispetto ai tradizionali conduttori in rame. Di ciò può trarre profitto anche la tecnologia dell'elaboratore per realizzare le interconnessioni del sistema.
G. GRASSO, G. SCALZO

Informatica e psicologia

Possono i concetti dell'informatica contribuire alla comprensione dei problemi della psicologia? L'Autore esprime in merito un suo originale punto di vista.
B. MELTZER

Ricerca documentaria e automazione

L'"information retrieval" è una delle più importanti aree di applicazione dell'elaboratore. L'articolo riassume la problematica relativa e fornisce interessanti esempi di realizzazioni in questo campo.
R. TESI

Diffusione dei piccoli sistemi gestionali nell'industria manufatturiera

L'evoluzione tecnologica mette gli strumenti dell'informatica a disposizione ormai anche dei piccoli utenti. L'articolo presenta i risultati di una indagine di mercato relativa alla piccola industria manufatturiera.
D. ROVINA

Direttore Responsabile
F. Filippazzi

Comitato di Redazione
C. Falcetti, P. Lupo,
M. Nobile, G. Occhini
G. Rapelli, F. Sala

Redazione
Honeywell Information Systems Italia
Centro di Ricerca e Progettazione
20010 Pregnana Milanese (Milano)

Stampa
tipolito Maggioni s.a.s.

Autorizzazione del Tribunale di Milano
n. 93 del 20 Marzo 1974

«Quaderni di Informatica» ospita articoli e contributi di varia provenienza. La responsabilità delle opinioni espresse rimane agli Autori.

La riproduzione di articoli della rivista, o di parte di essi, è consentita solo citando la fonte e l'autore.

ADA: un linguaggio per gli anni '80

GIAN LUIGI CASTELLI
Istituto di Cibernetica
Università degli Studi
Milano

1. Introduzione

Il presente lavoro ripercorre brevemente i passi che hanno determinato la progettazione e l'implementazione di ADA, il nuovo linguaggio di programmazione prescelto dal Dipartimento della Difesa (DoD) degli Stati Uniti quale elemento unificante nella progettazione di tutto il nuovo software che negli anni a venire sarà sviluppato per conto di esso.

Muovendo da alcune motivazioni specifiche che in ambito DoD hanno determinato la nascita di ADA, ne verrà ripercorsa la genesi, già di per sè piuttosto inusuale, e ne verranno tracciate le caratteristiche più interessanti, in modo da costituire un compendio per un primo approccio velatamente tecnico.

2. Le origini

Il Dipartimento della Difesa (DoD) degli Stati Uniti è l'ente governativo sotto il cui controllo ricadono le tre Armi (Esercito, Marina ed Aviazione) oltre ad un certo numero di sottoenti di minore consistenza, sia ad indirizzo terziario che tecnico.

Fin dagli inizi degli anni sessanta ha costituito uno dei principali, e più sofisticati, acquirenti di calcolatori, con un bilancio in cui le voci più sostanziose riguardavano, coerentemente con lo stato dell'arte dell'epoca, l'acquisto di nuovo hardware.

Tuttavia, già all'inizio degli anni settanta si era manifestata all'interno dell'ente, come del resto altrove seppure in maniera ridotta in proporzione alle dimensioni ed alle aspettative tecnologiche, una inversione di tendenza tra le diverse voci del bilancio.

Si stava verificando cioè una progressiva riduzione dei costi dell'hardware che associata al miglioramento generale delle prestazioni andava determinando nuove aspettative e la richiesta di nuove applicazioni altamente sofisticate quali il controllo automatico di sistemi balistici e di sistemi di protezione radar, la creazione e la gestione di grandi basi di dati oltre al normale sviluppo di programmi di calcolo scientifico e di sistemi gestionali.

A complicare ulteriormente le cose veniva l'aspetto altamente distribuito di questi sistemi; molti di essi costituivano infatti gli elementi dislocati su aerei, missili, navi e sommergibili di sistemi di controllo più vasti. Addirittura venne coniato un nome specifico per denotare questi sistemi: "embedded computer systems".

L'atteggiamento naturale determinato da un ambiente dalle esigenze così diverse fu quindi quello di sviluppare una grande quantità di software strettamente legato ad un gran numero di calcolatori differenti per produttore (la Marina sembrava infatti prediligere macchine Sperry Univac, mentre l'Aviazione utilizzava calcolatori Honeywell), per dimensione (dai grandi general purpose ai piccoli sistemi di controllo dedicato), per configurazione (dai normali sistemi con dischi, nastri, stampanti e così via, ai sistemi connessi a periferiche specializzate quali dispositivi di controllo di rotta e sistemi radar).

Molti di questi sistemi, essendo dedicati, non disponevano neppure di adeguati sistemi di sviluppo, il che richiedeva a sua volta lo sviluppo di nuovo software di emulazione e di strumenti cross su opportuni sistemi ospite.

Il risultato infausto fu pertanto quello di sviluppare non solo il software necessario, ma molto spesso anche quello superfluo, al punto che furono prodotti compilatori per qualche dialetto di linguaggi a loro volta a scarsissima diffusione e attraverso i quali furono realizzate solo poche applicazioni.

La comunicazione tra i gruppi risultava pertanto estremamente difficoltosa, e ciò si traduceva inevitabilmente in una moltiplicazione degli sforzi.

L'effetto finale di questa congiuntura fu che nei primi anni settanta, come risultò da un censimento successivo, erano in uso all'interno della comunità dei programmatori DoD più di 450 linguaggi di programmazione, che il costo della manutenzione del software esistente assorbiva la parte prepon-

derante delle spese relative al software, mentre globalmente queste superavano di larga misura quelle relative all'hardware.
Una previsione di spesa relativa all'anno 73/74 mostrava come limiti inferiori di spesa relativi al software importi oscillanti tra i 3 e i 3.5 miliardi di dollari. L'esigenza di un cambiamento che garantisse il contenimento delle spese unitamente a precise esigenze tecniche introdussero un'aspettativa senza precedenti nei confronti di una soluzione globale.
Un'analisi del software tipicamente sviluppato ed utilizzato in ambiente DoD dimostrò come mediamente i programmi impiegati fossero di grandi dimensioni, vale a dire intorno alle 100.000 linee di codice, e come la vita media degli stessi fosse ragionevolmente lunga, con un massimo di 15 anni di impiego.
Gli elevati costi della manutenzione si sommavano perciò alla difficoltà di trasporto dei programmi tra calcolatori che mutavano, in virtù del progresso tecnologico, assai più rapidamente della media delle applicazioni.
Fu quindi riconosciuto che la mancanza di un linguaggio che garantisse un elevato livello di standardizzazione e che costituisse uno strumento unificante in virtù della completezza della definizione, sebbene non fosse la panacea risolutiva, avrebbe costituito un primo significativo passo.
Ulteriori aspetti positivi di un linguaggio unificante furono riconosciuti in:
- riduzione dei costi di addestramento del personale
- portabilità dei programmi tra hardware altrimenti incompatibili
- focalizzazione degli sforzi sull'obbiettivo reale, senza dispersioni di energie tra obbiettivi e strumenti
- elevato grado di comunicazione tra differenti gruppi di lavoro
- concentrazione dei fondi su programmi comuni
- definizione di un ambiente di programmazione standard

3. La Genesi

Il risultato degli studi condotti da parte degli uffici del DoD trovavano quindi conferma nei problemi e nelle aspettative delle tre Armi.
Queste stesse diventarono, nel 1974, le dirette promotrici dello sforzo di ricerca verso un linguaggio unificante.
Le proposte raccolte furono accolte all'inizio dell'anno successivo dal Sottosegretariato della Difesa, della Ricerca e dell'Ingegneria del DoD (USDRE). Il prodotto iniziale dell'attività di tale ente fu una prima bozza di un programma comune di lavoro, ma soprattutto la misura drastica ma necessaria consistente in un taglio di tutti i fondi associati alla ricerca ed allo sviluppo di nuovi linguaggi di programmazione.
Questa bozza di piano di lavoro vedeva coinvolto un gruppo di lavoro costituito da rappresentanti delle tre Armi. Ben presto però ci si rese conto che un allargamento del gruppo a persone di ambienti diversi e con differenti esperienze avrebbe potuto introdurre nuove problematiche e dunque estendere l'applicabilità del nuovo linguaggio ad istanze diverse.
In ambito nazionale, la cooperazione venne estesa alla NASA ed ad altre agenzie governative di minore importanza.
Un aspetto molto interessante dell'approccio seguito fu rappresentato dall'estensione del gruppo di lavoro ad enti militari alleati con particolare riguardo al Ministero della Difesa britannico.
Lo scopo iniziale del gruppo di lavoro eterogeneo che era stato così costituito fu quello di selezionare una lista di pochi linguaggi significativi che potessero venire impiegati nello sviluppo di nuovo software, riducendo la dispersione degli sforzi.
In secondo luogo, un'analisi comparata approfondita di questi linguaggi sembrava poter evitare lo sviluppo di un nuovo linguaggio.
Un'ulteriore attività del gruppo venne riconosciuta nel coordinamento delle ricerche, disgiunta altresì da qualunque sforzo di progettazione o di implementazione.
Dopo circa tre mesi di attività, e cioè nella primavera del 1975, fu prodotto un elenco di soli sette linguaggi temporaneamente accettabili, ma ritenuti sostanzialmente non idonei per un uso prolungato a causa della loro scarsa generalità ed adeguatezza.
Alcuni di questi linguaggi vennero scelti principalmente per l'elevato grado di standardizzazione da essi raggiunto, benché attualmente inadeguati e limitati nel raggio delle applicazioni, quali l'ANSI Fortran e l'ANSI Cobol, altri per l'applicabilità a problemi di tempo reale e per il loro vasto uso all'interno del DoD, sebbene largamente sconosciuti in altre organizzazioni, quali il Tacpol, il CMS-2, lo SPL-1, il J3 Jovial ed il J73 Jovial. L'incapacità dei linguaggi selezionati di soddisfare globalmente le esigenze che erano state individuate diedero il via alla vera fase di sviluppo del progetto, consistente nella definizione organica e precisa dei requisiti del nuovo linguaggio.
Le fasi di sviluppo attraverso cui si giunse alla definizione del linguaggio rappresentarono un modo davvero inusuale di condurre lo sviluppo di un progetto software, sia per il numero di persone ed organizzazioni implicate, sia per l'abbondante numero di revisioni e di rielaborazioni successive dei requisiti, che tesero a definire un linguaggio universale.
La prima fase venne chiamata STRAWMAN ("uomo di paglia"), per denotare la fragilità, e talvolta l'inconsistenza, che i requisiti in esso contenuti possedevano, anche in virtù di esigenze solo genericamente specificate.
Certamente assodati erano invece i concetti fondamentali della programmazione strutturata che uscivano da più di un lustro di intensa applicazione negli ambienti più disparati e che si erano riflessi sui progettisti dei linguaggi di scuola europea quali Dahl (Simula 67), Van Wingardeen (Algol 68), Brinch Hansen (Concurrent Pascal), Ichbiah (Lis) e probabilmente il più noto di tutti, Wirth (Pa-

scal, Modula e Modula-2), ed avevano determinato le forme dei linguaggi più innovativi ed accreditati.

Le indicazioni fornite avevano dunque solo carattere generale e si limitavano a sottolineare l'esigenza di semplicità, efficienza ed affidabilità che il nuovo linguaggio avrebbe dovuto rispecchiare. Il documento prodotto non ebbe una circolazione molto vasta, limitata in pratica al solo ambiente militare.

Una bozza più precisa venne stilata nell'autunno del 1975 sulla base delle reazioni al documento ed attraverso un questionario a vasta diffusione sulle ipotetiche caratteristiche di un linguaggio "ideale".

L'integrazione delle informazioni in un documento determinò l'inizio della seconda fase di definizione, che venne denominata WOODENMAN ("uomo di legno").

In questa fase venne coinvolto nel lavoro di revisione e di suggerimento un ambiente notevolmente più vasto: furono distribuite copie sia in ambiente accademico che industriale ed anche fuori dagli Stati Uniti in ambito CEE.

La mole dei suggerimenti, delle richieste di modifiche e delle puntualizzazioni produssero un documento più esteso nel quale a fronte di ogni specifica comparivano le motivazioni che l'avevano determinata.

L'approvazione dei contenuti da parte del comitato di controllo fece iniziare l'attività della terza fase che venne chiamata TINMAN ("uomo di latta").

Questa fase ebbe carattere sostanzialmente transitorio e di riflessione, al termine della quale venne rilasciato un documento che si configurava come l'insieme delle specifiche funzionali, dunque senza indicazioni nè di forma nè implementative, del linguaggio "ideale" IRONMAN ("uomo di ferro").

Si era ormai giunti al luglio 1977. A fronte delle specifiche del documento IRONMAN, considerate sufficientemente stabili, venne condotta una ulteriore inchiesta sull'adeguatezza dei linguaggi esistenti e sulla loro copertura dei requisiti.

L'analisi assunse proporzioni mondiali, coinvolgendo sedici organizzazioni diverse, in un vasto spettro di ambienti: accademici, industriali e governativi.

L'analisi riguardò ventitre linguaggi. Di nuovo l'esito fu negativo poiché nessun linguaggio era in grado di coprire completamente i requisiti di IRONMAN. Tuttavia l'indicazione di massima fu precisa: tre dei linguaggi esaminati potevano comunque costituire una base di sviluppo adeguata, seppure con motivazioni diverse. Due di questi, di scuola europea, furono considerati per la pulizia di disegno: Pascal e Algol 68; il terzo per la completezza delle caratteristiche: PL/1.

Lo sforzo congiunto di tre anni di lavoro venne allora congelato nel documento di specifiche conclusivo: STEELMAN ("uomo d'acciaio").

In esso si richiese che il linguaggio fosse sviluppato sulla base delle indicazioni qui sinteticamente elencate:

- basato su uno dei tre linguaggi approvati come punto di partenza
- fortemente tipizzato
- altamente strutturato
- dotato di strutture per la gestione del parallelismo
- dotato di strutture per la gestione delle condizioni eccezionali
- in grado di definire e gestire la precisione dell'aritmetica
- in grado di gestire compilazioni separate con librerie di programmi per il controllo della compatibilità.

Per la realizzazione del linguaggio venne bandito un concorso aperto a chiunque desiderasse partecipare ed anonimo sino alla selezione finale.

Il clima di elevata competitività non mancò di suscitare proteste le più disparate da parte degli ambienti più diversi.

Oggettivamente la moltiplicazione degli sforzi fu notevole poiché parteciparono al concorso quindici gruppi diversi con altrettante proposte di linguaggi.

D'altra parte l'approccio seguito garantì, con ogni probabilità, la formazione di gruppi di lavoro di dimensione più contenuta e con una visione certamente più omogenea del progetto. Due condizioni, queste, essenziali a causa della ristrettezza dei tempi di sviluppo, e certamente garanzia di maggiore economicità per il DoD.

Da un punto di vista tecnico, la maggioranza dei linguaggi proposti erano basati su Pascal e tra questi i quattro finalisti. Essi furono denominati in codice RED, BLUE, GREEN e YELLOW e furono disegnati rispettivamente da Softech, SRI International, CII-Honeywell Bull ed Intermetrics.

L'attività relativa allo sviluppo dei quattro linguaggi finalisti terminò la sua prima fase nel febbraio del 1978.

I manuali dei linguaggi vennero pubblicati su una rivista del settore a larga diffusione, così da renderne note le caratteristiche a chiunque fosse interessato.

La revisione dei manuali venne intrapresa da un gruppo di quattrocento persone e costituì un momento di raffinamento.

Nel corso del 1979 venne conclusa la fase di progettazione e venne decretata l'accettazione da parte del DoD del linguaggio GREEN, realizzato dalla consociata francese della Honeywell, la CII-Honeywell Bull, sotto la direzione di J. Ichbiah.

Come europeo fu il gruppo di progetto, così il nome prescelto: Ada, in onore di Ada Augusta contessa di Lovelace, nonché figlia di Byron, amica e collaboratrice di Charles Babbage, di cui programmava le macchine da calcolo ed i telai controllati da schede perforate (fig. 1).

Nel corso del 1980 è iniziato lo sviluppo dei compilatori e lo studio per un ambiente portabile di programmazione associato ad Ada, anch'esso sotto l'egida del DoD, e denominato APSE (Ada Programming System Environment).

Fig. 1 - Ritratto di Ada Augusta contessa di Lovelace, figlia di Byron, che programmò le pionieristiche macchine da calcolo a schede perforate di Charles Babbage. In suo onore è stato scelto il nome del nuovo linguaggio, anche per ricordare l'origine europea del medesimo. Il linguaggio è stato infatti sviluppato dalla CII-Honeywell Bull, consociata francese della Honeywell.

4. Aspetti tecnici

I criteri generali che hanno guidato la definizione delle specifiche di Ada possono venir riassunti nei seguenti concetti essenziali:

- Forte tipizzazione: Ada contiene le indicazioni che sono state fornite dai linguaggi di programmazione degli anni settanta. Pertanto ogni oggetto deve essere associato esplicitamente ad un tipo che definisce la struttura dell'oggetto e le operazioni legali che su di esso possono venir effettuate. In tal modo il compilatore è in grado di realizzare controlli estensivi con una ricca diagnostica al momento della compilazione.
  L'insieme di alcuni tipi elementari predefiniti dal linguaggio, come gli interi, i reali, i caratteri, i booleani e gli enumerativi, uniti ad alcuni costruttori di tipo, come i record o gli array, permettono all'utente di definire le strutture di dati che più ritiene adatte alla risoluzione di un problema.
  Alcune ambiguità presenti in alcuni linguaggi, come Pascal, sul concetto di compatibilità di tipo sono state rimosse.
- Tipi di dati astratti: la possibilità di nascondere dettagli implementativi di un modulo di programma, permettendo di distinguere il momento di implementazione dal momento di uso, è stata adottata in Ada.
  Anche questa caratteristica era presente in alcuni dei linguaggi esistenti a partire dal Simula 67, attraverso il Concurrent Pascal, fino ad altri linguaggi sperimentali come CLU, Mesa o Alphard.
  È quindi possibile definire oggetti che sono composti da strutture di dati e da procedure (o funzioni) che operano su queste strutture. Le operazioni così definite sono le sole legali. È possibile cioè implementare, ad esempio, uno stack come un vettore, le operazioni di push e di pop come incrementi o decrementi dell'indice del vettore, senza che tutto ciò sia visibile all'esterno. Chi usa il tipo astratto vede solo un oggetto organizzato a stack e due operazioni su di esso: push e pop.
- Gestione delle eccezioni: Ada definisce uno strumento molto efficace per il controllo da parte del programma delle condizioni eccezionali. Attualmente, per molti linguaggi, il verificarsi di una condizione di errore durante l'esecuzione, sia per cause interne che esterne, determina l'arresto della esecuzione. Così il tentativo di una divisione per zero od un errore di lettura dei dati arresta l'esecuzione di un programma Pascal, mentre un programma Ada può gestire al suo interno le azioni più opportune per la continuazione della esecuzione.

- Multiprogrammazione: molte delle applicazioni per cui Ada è stato pensato sono a processi concorrenti o multitask. Sono stati perciò introdotti strumenti che facilitino il lavoro del programmatore, garantendo ad un tempo un elevato grado di correttezza al momento della compilazione, riducendo così la possibilità di introdurre errori dipendenti dal tempo, di difficile individuazione.
  La soluzione introdotta è più innovativa di quelle presenti in altri linguaggi come il Concurrent Pascal od il Modula, dove il concetto ispiratore era il monitor. Ada si fonda, a questo proposito, su lavori recenti di P. Brinch Hansen e di C.A.R. Hoare, fornendo una prima implementazione del concetto di rendezvous.
- Programmazione in grande: Ada è stato realizzato per la produzione di software in ambienti di grandi dimensioni per grandi progetti, con conseguente parcellizzazione del lavoro. Risultano perciò essenziali i requisiti che introducono una semplice ed efficace divisione del lavoro. La programmazione top-down risulta facilitata dalla possibilità di costruire librerie comuni di programmi, dalla possibilità di incapsulare algoritmi specifici in moduli garantendo accessi controllati e selettivi a questi, dalla possibilità di gestione flessibile delle compilazioni separate.
- Ordinamento gerarchico: è possibile specificare in Ada dei programmi che precedentemente compilati costituiscano delle basi di dati per le compilazioni successive. Sono garantite in questo modo le interfacce di comunicazione tra moduli compilati separatamente, la compatibilità di oggetti appartenenti a moduli diversi e l'instanziabilità di oggetti parametrici.

5. Struttura di programma

Un programma Ada può essere costituito da una sola unità di compilazione, che contenga eventualmente sottoprogrammi interni, oppure essere costruito su più moduli compilati separatamente. Questi moduli costituiscono la libreria del programma.

I moduli attraverso i quali un programma può essere costruito sono di tre tipi fondamentali: i sottoprogrammi, i package ed i task.

Un sottoprogramma è una procedura o una funzione all'interno di una unità di compilazione. Questo è anche il concetto che più facilmente si riscontra anche in altri linguaggi.

Un sottoprogramma può avere parametri di tre tipi: costanti, cioè a sola lettura, per valore e per indirizzo. Questi sono denotati da tre prefissi dichiarativi: in, out e in out.

Per evitare effetti collaterali le funzioni possono avere solo parametri di tipo in.

Un package permette di specificare insiemi di oggetti correlati tra loro da qualche relazione logica.

All'interno di un package sono ben distinte la fase di definizione e la fase di implementazione degli oggetti che vi sono contenuti.

Un package può assumere forme diverse: può essere una raccolta di dichiarative di dati condivisibili da altri package, oppure una raccolta di dati e di sottoprogrammi, oppure può essere un tipo di dato astratto che protegge gli oggetti che vi sono contenuti rendendoli visibili in modo selettivo e controllato. Il mondo esterno può accedere agli oggetti di un package aprendo esplicitamente una finestra attraverso cui vedere solo una parte del mondo interno.

Un task è uno strumento per la gestione del parallelismo. La sincronizzazione e la comunicazione tra diversi task avvengono attraverso "appuntamenti" tra un task che effettua una chiamata ed uno che la accetta.

Sintatticamente i task non differiscono dai package. Tuttavia, mentre quest'ultimi contengono solo oggetti che possono essere attivati, e sono dunque passivi, i task sono degli elementi attivi.

Perché un task possa comunicare con un altro occorre che il ricevente dichiari quali sono gli ingressi in cui devono avvenire gli appuntamenti.

Il task che effettua la chiamata viene sospeso finché il ricevente non giunge all'appuntamento. Terminato l'incontro i task procedono in maniera indipendente.

6. I tipi di dati

Come si è precedentemente accennato, ad ogni oggetto è associato un tipo. I tipi possono essere predefiniti o specificati dall'utente secondo le sue esigenze attraverso opportuni costruttori. La complessità di nidificazione strutturale è arbitraria.

I tipi di dati semplici sono: INTEGER, BOOLEAN, CHARACTER, FLOAT e gli oggetti definiti per enumerazione.

I costruttori di tipo, che definiscono invece le regole con cui aggregare tipi semplici o strutture precedentemente definite, sono:

- ARRAY: una struttura di dati costituita da elementi tutti dello stesso tipo, ordinati in forma tabellare ad una o più dimensioni ed accessibili direttamente per posizione.
- RECORD: una struttura di dati costituita da elementi di natura anche diversa, detti campi, ed accessibili attraverso i nomi dei campi. È possibile anche definire varianti per accedere ad elementi di insiemi disgiunti di informazioni, con risparmio di memoria e chiarezza concettuale.
- POINTER: i puntatori sono oggetti che permettono la creazione dinamica di oggetti del tipo associato al puntatore, in una area di memoria chiamata "heap". Gli oggetti associati al puntatore vengono allocati mediante la procedura predefinita NEW.

Contrariamente ad altri linguaggi, non sono presenti in Ada il costruttore SET ed il costruttore FILE.

Il primo è sostituito dalla più semplice ed efficiente implementazione degli array di bit, il secondo viene incapsulato in package dedicati all'I/O.

7. Le strutture di controllo

Ormai non resta molto da dire relativamente alle strutture di controllo.

Infatti sono presenti in Ada i classici costrutti di sequenza, iterazione e selezione, con sintassi leggermente diverse da quelle esistenti.

Esistono perciò tre tipi di costrutti di selezione:

- IF espressione THEN statement END IF
- IF espressione THEN statement1 ELSE statement2 END IF
- IF espressione THEN statement1 ELSIF espressione2 THEN statement2 END IF

e un costrutto di selezione multipla:

- CASE espressione IS
  WHEN label1 → statement1
  WHEN label2 → statement2
  WHEN label ....
  END CASE

I costrutti di iterazione sono tre:

- FOR variable IN range LOOP
  statement
  END LOOP
- WHILE espressione LOOP
  statement
  END LOOP
- LOOP
  statement
  END LOOP

Tutti i costrutti iterativi possono essere terminati da un verbo EXIT.

Comune a tutte le strutture di controllo è la presenza dei terminatori di costrutto che eliminano la necessità delle coppie BEGIN END o di altre forme di terminazione la cui parità spessa causa problemi.

8. Bibliografia

A.A. VARI - Rationale for the Design of the ADA Programming Language - *Sigplan Notices of the ACM* - vol. 14, n. 6, giugno 1979, parte B.

C.A.R. HOARE - Communicating Sequential Processes - *Communications of the ACM* - vol. 21, n. 8, agosto 1978.

P. BRINCH HANSEN - Distributed Processes: a Concurrent Programming Concept - *Communications of the ACM*, vol. 21, n. 11, novembre 1978.

J.G.P. BARNES - An Overview of ADA - *Software Practice and Experience* - vol. 10, n. 11, novembre 1980.

Reference manual for the ADA programming language - proposed standard document - U.S. DoD - luglio 1980.

Supercalcolatori per applicazioni speciali

JAMES W. BOWRA
Honeywell
Marine Systems Center
Seattle, USA

1. Introduzione

L'elaboratore che gestisce un moderno sistema "sonar" (radar subacqueo) deve essere in grado di eseguire 10 milioni di moltiplicazioni e addizioni al secondo. Anche se questo valore può sembrare grande, esso risulta piccolo se paragonato coi 100 milioni o più di moltiplicazioni e addizioni per secondo richiesti dalla prossima generazione di "sonar". Questi valori sono al di là della capacità di un singolo elaboratore, anche impiegando i più avanzati e veloci circuiti integrati. Occorre quindi passare ad architetture multi-processore. Questo articolo presenta una architettura di tale tipo, basata su un "bus a matrice" (*) che consente di superare le limitazioni di velocità legate alla interconnessione dei moduli costituenti il sistema, rendendo possibile la realizzazione di elaboratori aventi 32 o più processori.

(*) Questo concetto è stato proposto originariamente da Bill Wren, del Marine Systems Center della Honeywell.

2. L'elaborazione digitale di segnali

Nella elaborazione digitale di segnali, un segnale elettrico viene convertito in una sequenza di numeri (per esempio, una serie di valori rappresentanti le ampiezze istantanee), mentre altri parametri caratteristici del segnale sono stimati mediante manipolazioni ripetitive di tali numeri. I parametri stimati possono essere semplici, quale le frequenze componenti il linguaggio umano, o complicati, come lo spettro di frequenza di una specifica parola.

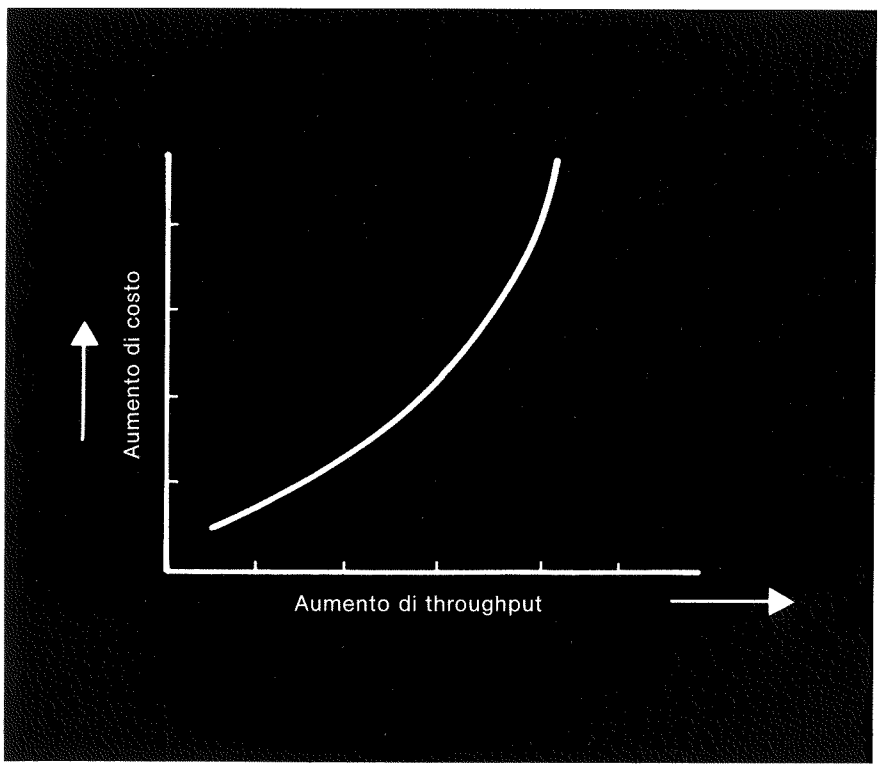


Fig. 1 - Il costo di un elaboratore digitale aumenta non linearmente con la sua capacità di elaborazione ("throughput"). A partire da un certo punto, inoltre, le prestazioni non possono più essere aumentate se non ricorrendo a nuove architetture del sistema.

L'elaborazione dei segnali è usata nei sistemi di identificazione di oggetti o bersagli mediante radar, raggi infrarossi o onde di altra natura, tipo quelle acustiche propagantisi nell'acqua, come nel caso del "sonar" cui si fa riferimento in questo articolo.

La relazione tra le prestazioni di un elaboratore digitale di segnali costruito con circuiti integrati standard ed il costo relativo, risulta fortemente non lineare; al crescere delle prestazioni (throughput), il costo aumenta drammaticamente (fig. 1). Oltre un certo punto, determinato dalla tecnologia dei circuiti integrati, diventa del tutto impraticabile spingere ulteriormente le prestazioni. Anche se i circuiti integrati sono progettati ad hoc (circuiti "custom"), presenteranno pur sempre limitazioni delle loro prestazioni, mentre il costo del progetto sarà in questo caso molto più alto. Al momento, i requisiti di elaborazione di molti sistemi di visualizzazione in tempo reale non possono essere soddisfatti mediante elaboratori digitali programmabili; occorre usare tecniche analogiche oppure sistemi digitali con hardware non programmabile.

Risulta perciò molto interessante l'idea di aumentare la velocità di elaborazione collegando numerosi processori tra di loro. L'aumento del numero di processori aggiunge però oneri addizionali alla gestione del sistema, e precisamente il tempo che l'elaboratore deve spendere per sincronizzare i processori. Come risultato, aumentando il numero di processori oltre un certo punto, le prestazioni complessive del sistema diminuiscono (fig. 2). Il valore di questo punto dipende da molti fattori, incluso il tipo di calcoli da eseguire. Con elaboratori general purpose questo limite si raggiunge abbastanza rapidamente; ad es. usando il MULTICS, il numero ottimale di elaboratori interconnessi è inferiore a dieci. La natura strutturata dell'elaborazione digitale di segnali consente, invece, di interconnettere efficientemente un numero di processori considerevolmente maggiore ove tale caratteristica venga opportunamente sfruttata.

L'elaborazione digitale di segnali risulta infatti altamente ripetitiva e richiede soltanto pochi tipi di operazioni matematiche. Il segnale di ingresso, proveniente da un sensore, viene convertito in una sequenza di numeri, per esempio, una serie di valori rappresentanti la tensione in un punto di un circuito elettrico, misurata ad intervalli regolari di tempo:

$$V_1, V_2, V_3 \cdots V_n$$

Questa è chiamata una serie temporale. Una tipica operazione matematica nell'elaborazione dei segnali è la somma di prodotti. I dati forniti da ciascun punto sono moltiplicati per una costante ("peso"), e questi prodotti sono sommati tra loro:

$$X_t = \sum_{i=0}^{N} C_i V_{t-i}$$

In questo modo, viene generata una nuova serie temporale X_t a partire da quella precedente. Si noti che la sequenza di calcolo è indipendente dai valori dei dati (V_{t-i}).

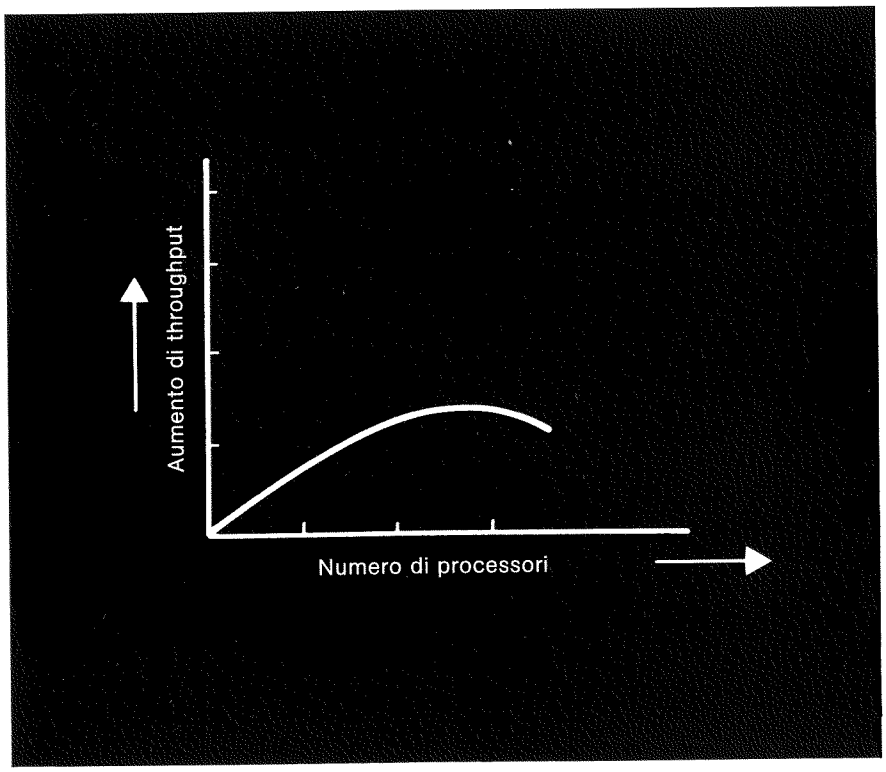


Fig. 2 - Le prestazioni di un sistema multi-processore aumentano col numero dei processori fino ad un certo punto, dopo il quale l'aggiunta di altri processori degrada le prestazioni.

Un esempio di questo tipo di elaborazione è rappresentato da un filtro "passa-basso" (FIR = finite pulse filter). Un semplice caso di FIR richiede la media di due tensioni in una serie temporale. In questo caso si ha cioè N = 2, mentre C = 1/2:

$$X_t = \frac{V_t}{2} + \frac{V_{t-1}}{2}$$

Scopo di questa elaborazione del segnale è di bloccare le frequenze più alte del segnale stesso, che sono spesso semplicemente dei disturbi. La fig. 3 mostra come il segnale X_t risulti più "liscio" di quello originale V_t.

La tabella di fig. 4 elenca alcuni dei calcoli più comuni nella elaborazione dei segnali. Uno dei più importanti è la trasformata di Fourier. Un segnale periodico può essere scomposto per mezzo della trasformata di Fourier in un numero infinito di componenti sinusoidali. Le frequenze di queste componenti sono le armoniche di alcune frequenze base. Se le frequenze componenti il segnale cadono soltanto in una banda ristretta, come avviene in un "sonar" dopo la pre-elaborazione del segnale, la somma risulta costituita da un numero finito anziché infinito di termini. La trasformata "discreta" di Fourier identifica le frequenze che compongono un segnale rappresentato da una serie temporale.

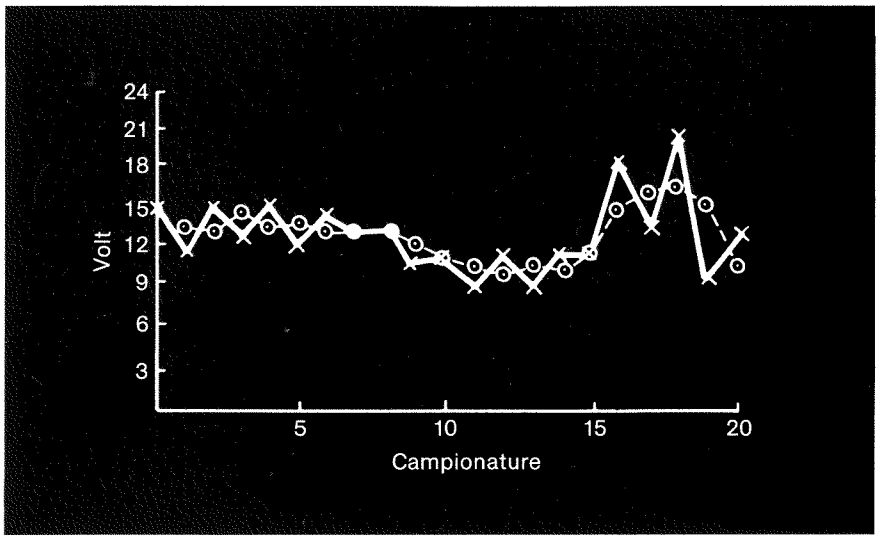


Fig. 3 - Un filtro passa basso riduce le componenti ad alta frequenza del segnale, rendendolo più "liscio".

Funzione	Operazione
Spostamento di banda	$Y_t = X_t \exp\left(\frac{j2\pi t}{N}\right)$
Filtro digitale	$Y_t = \sum_{i=0}^{N-1} C_i X_{t-i}$
Finestra	$Y_t = C_{t(\mathrm{mod})N} X_t$
Trasformata rapida di Fourier (FFT)	$Y_t = \sum_{k=0}^{N-1} X_k \exp\left(\frac{-2\pi jtk}{N}\right)$
Ampiezza	$Y_t = \sqrt{X_t^2 + X_{t-i}^2}$
Media temporale	$Y_t = \alpha X_t + (1-\alpha) X_{t-N}$

Fig. 4 - Tabella delle più comuni funzioni che si incontrano nell'elaborazione dei segnali.

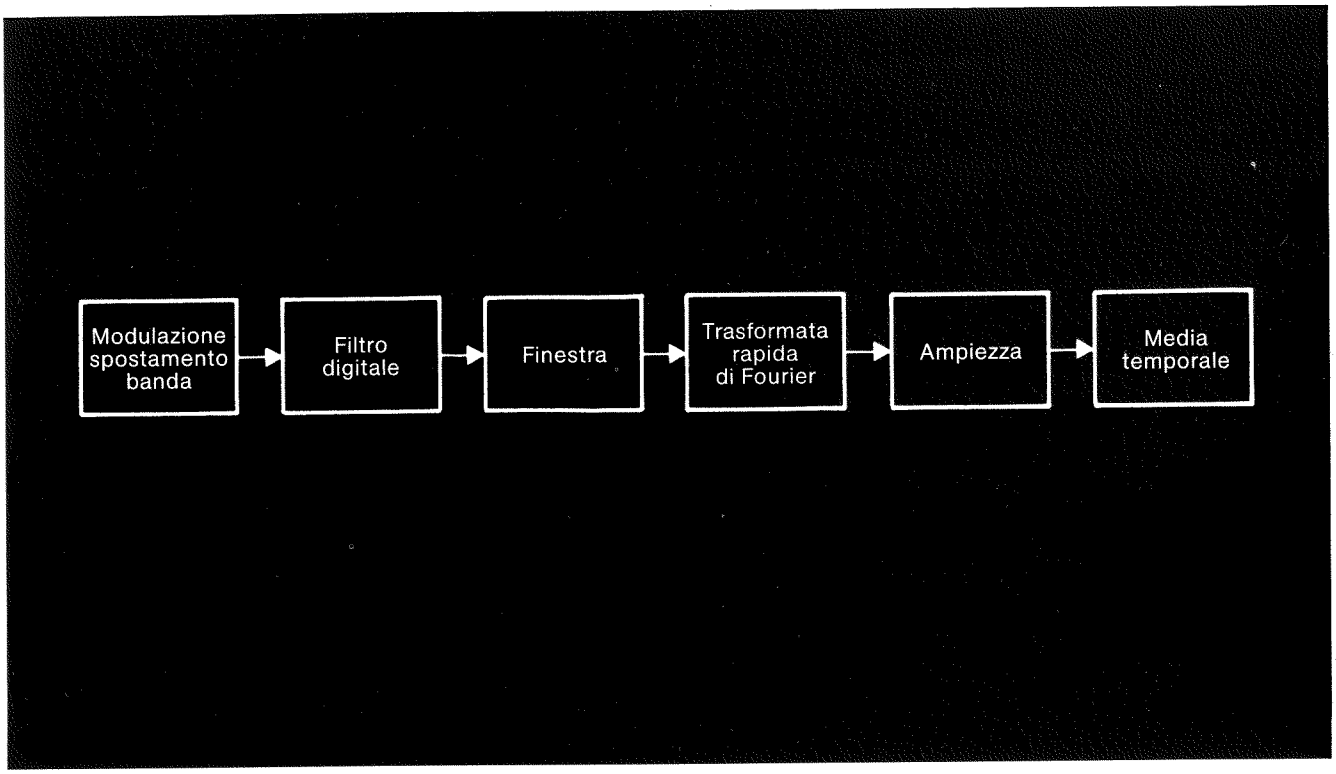


Fig. 5 - Le funzioni richieste nell'analisi spettrale (che determina le frequenze componenti un segnale) e la sequenza in cui esse sono trattate, illustrano la natura a "pipeline" delle elaborazioni sui segnali.

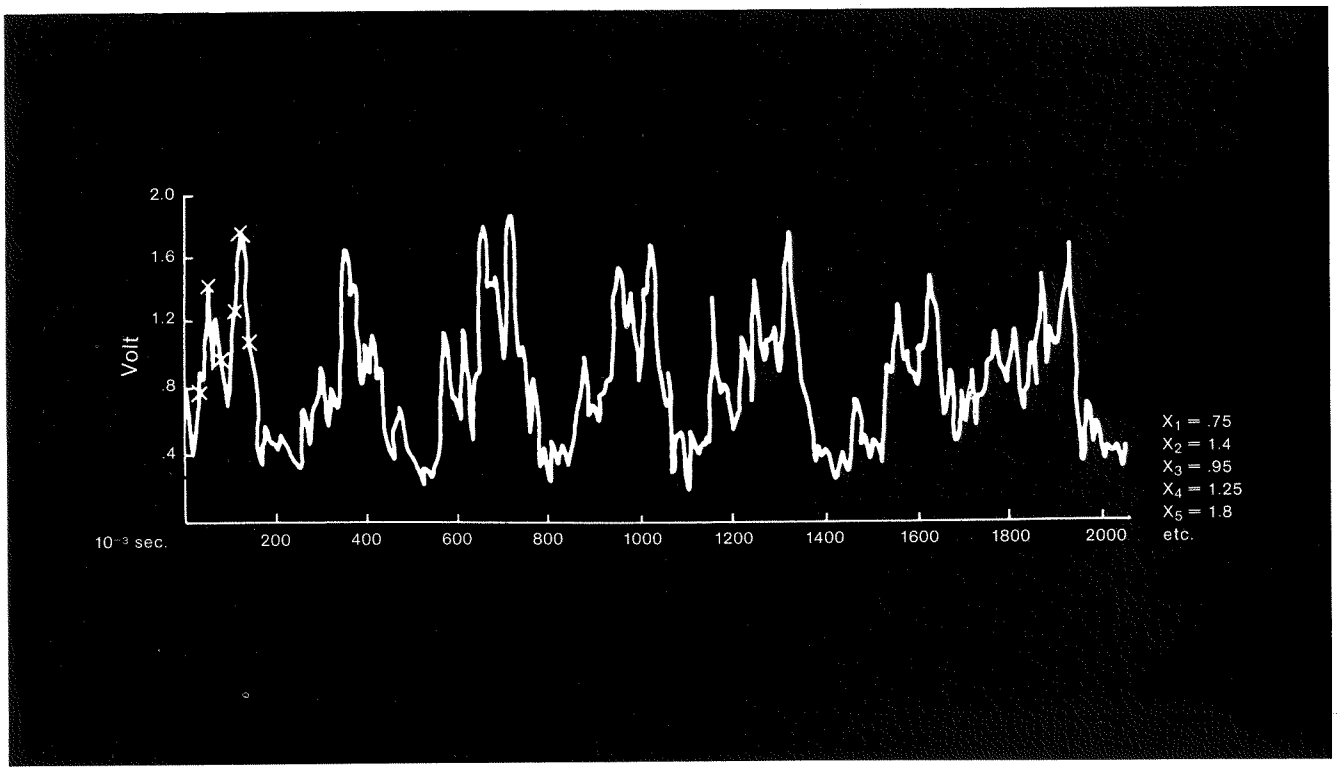


Fig. 6 - L'ingresso analogico ad un sistema di analisi spettrale viene campionato periodicamente, convertito in valori numerici e poi elaborato.

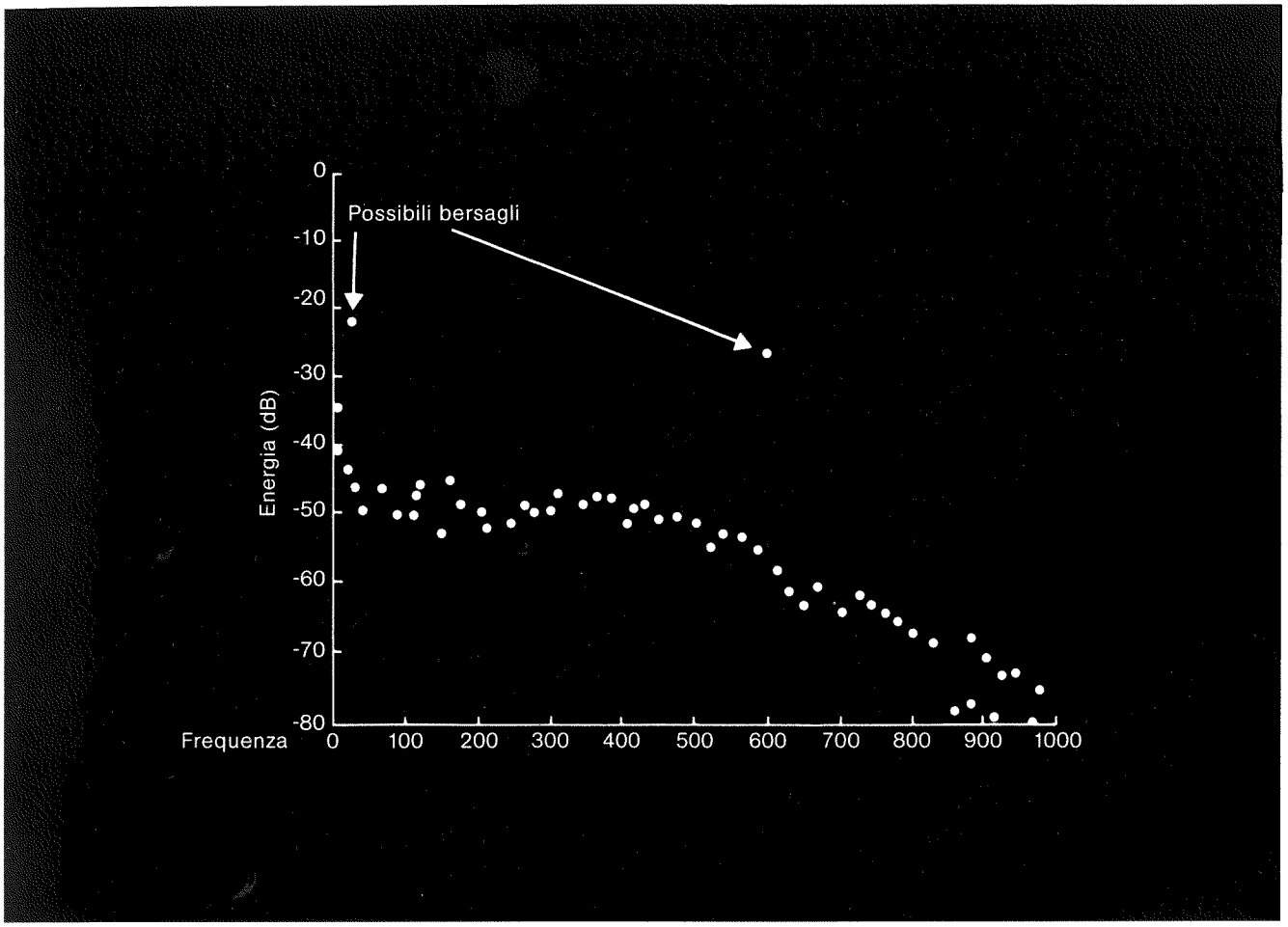


Fig. 7 - Le frequenze componenti un segnale "sonar" vengono determinate mediante la trasformata rapida di Fourier (FFT). Picchi nei valori di queste componenti possono identificare un eventuale bersaglio.

Le frequenze componenti un segnale "sonar" rinviato da un ostacolo, rivelano non solo la presenza ma anche le caratteristiche di quest'ultimo.

Esiste anche una trasformata rapida di Fourier (FFT = Fast Fourier Transform), che è una tecnica per raggruppare i termini della sommatoria riducendo l'ammontare di tempo di elaborazione occorrente per l'esecuzione di calcoli.

Le fig. 5, 6 e 7 illustrano rispettivamente le funzioni richieste per effettuare l'analisi di uno spettro "sonar", la trasformazione del segnale analogico di ingresso in una serie temporale e le frequenze componenti il segnale, rivelate dalla applicazione della FFT. La comparsa tra quest'ultime di picchi indica la presenza di un bersaglio in immersione.

3. Calcolatori per l'elaborazione di segnali

Gli elaboratori di segnali sono necessariamente specializzati poiché quelli general purpose non sono in grado di fornire le velocità richieste. Per esempio, per eseguire una FFT a 1024 punti (N = 1024), un elaboratore general purpose quale un Honeywell Livello 6 richiede diversi secondi; invece un elaboratore speciale come l'HAP (Honeywell Array Processor), di cui parleremo in seguito, che è notevolmente più piccolo come dimensioni fisiche, può eseguire fino a 200 FFT per secondo. In un minicomputer, i passi richiesti per eseguire operazioni matematiche sono tipicamente svolti in sequenza. Per esempio, per calcolare una somma di prodotti un minicomputer esegue i passi seguenti:

1. Calcola l'indirizzo di C_i
2. Calcola l'indirizzo di V_{t-i}
3. Estrae V_{t-i} e C_i dalla memoria
4. Moltiplica V_{t-i} per C_i
5. Aggiunge il risultato alla somma parziale
6. Verifica se i = N
7. Salta a 1 se i ≠ N

Ciascuno di questi passi richiede l'estrazione di una o più istruzioni dalla memoria. Le istruzioni e i dati sono registrati nella stessa memoria. Tipicamente, si ha un piccolo, se non nullo, parallelismo operativo e la moltiplicazione viene eseguita mediante una successione di scorrimenti (= moltiplicazione per 2) e di somme nell'unità aritmetica.

Gli elaboratori digitali di segnali sono costruiti con unità funzionali molto veloci ed il progetto è orientato a rendere massimo il grado di

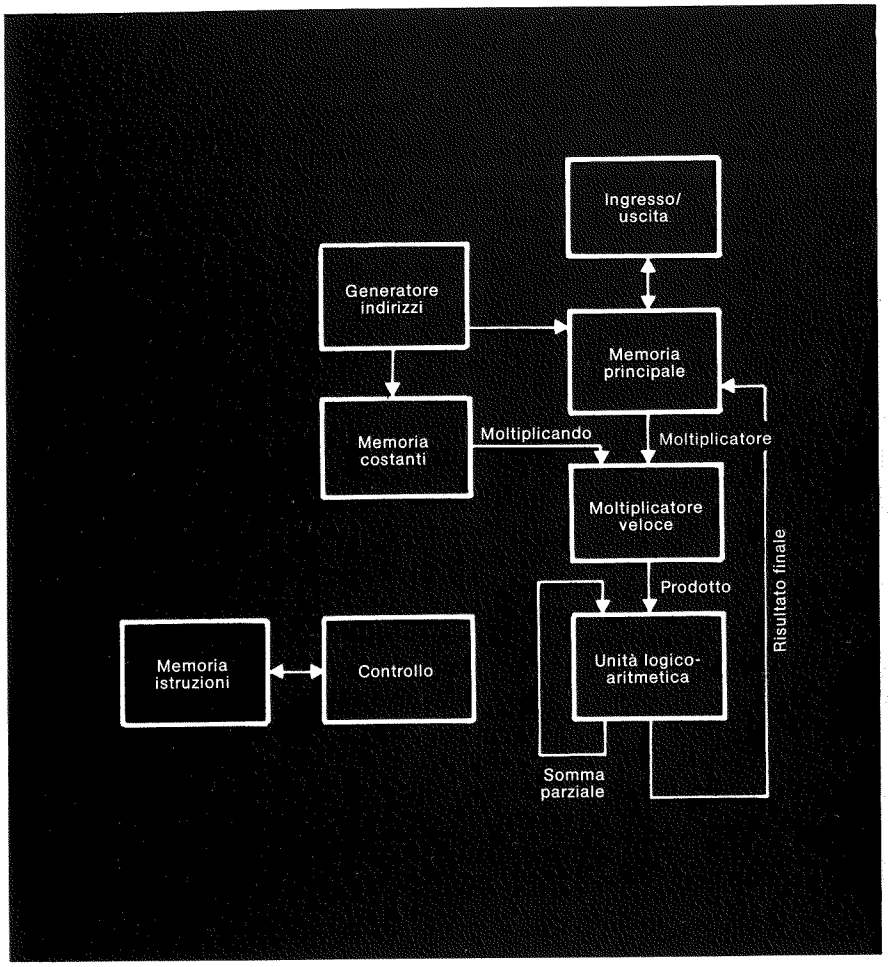


Fig. 8 - Questo schema semplificato del "Honeywell Array Processor" mostra i blocchi che compongono questo elaboratore, che è previsto possano funzionare tutti simultanemente.

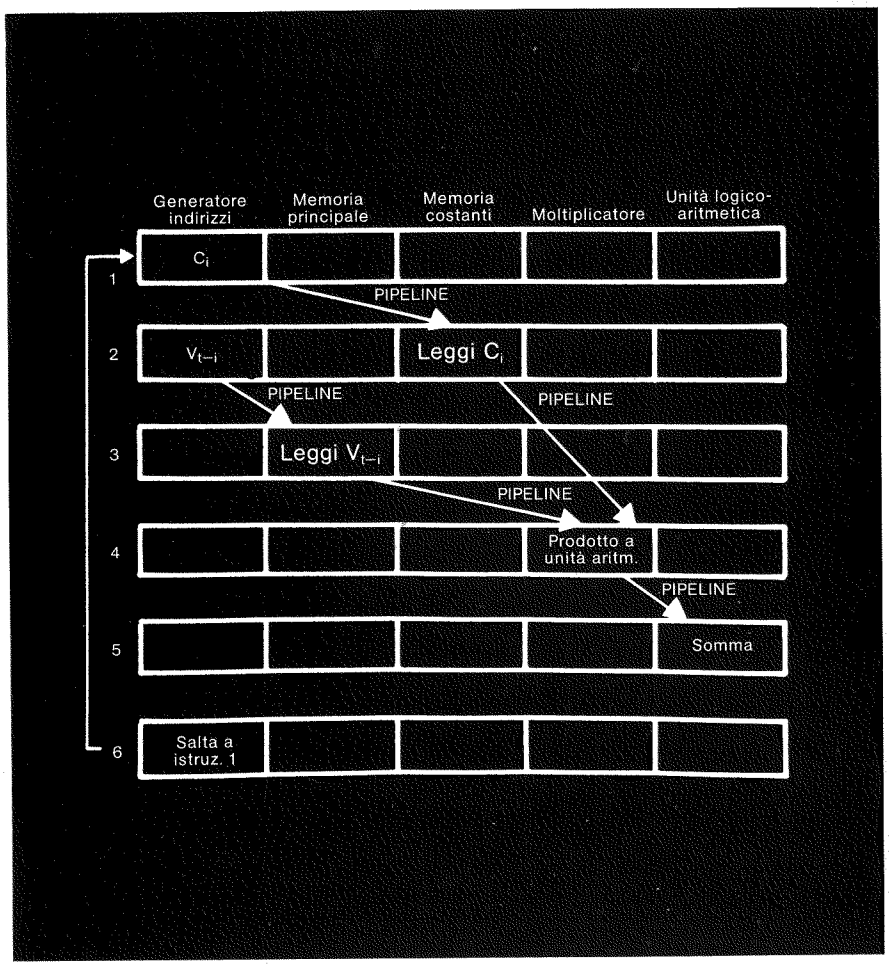


Fig. 9 - In un elaboratore di segnali, una istruzione controlla tutti gli elementi in ogni ciclo di istruzione ("parallelismo"). Tra un ciclo e il successivo, i dati passano da un elemento a quello adiacente ("pipeling").

"pipeling" (manipolazione sequenziale dei dati da parte di diverse unità, ciascuna delle quali esegue tipicamente un tipo solo di elaborazione) e di "parallelismo" (numero di unità di calcolo, quali le unità aritmetiche ed i moltiplicatori veloci, che operano simultaneamente). La differenza tra la manipolazione sequenziale dei dati in un minicomputer ed il "pipelining" in un elaboratore di segnali equivale alla differenza tra il caso di un operaio che monta da solo una macchina intera, un pezzo dietro l'altro, e quello degli operai di una linea di montaggio, ognuno dei quali esegue soltanto una specifica operazione.

Seguendo i passi con cui viene eseguita una somma di prodotti in un HAP (fig. 8) ci si rende conto di che cosa sia il parallelismo e il pipeling e di come essi influiscano sull'elaborazione complessiva. Gli elementi di calcolo di un HAP sono tutti controllati e sincronizzati da una singola istruzione. Quando la serie temporale arriva al convertitore analogico-digitale, essa viene registrata nella memoria principale. I coefficienti sono registrati in una tabella a parte, per consentire un accesso più rapido ad essi. Il generatore di indirizzi calcola gli indirizzi di memoria sia della serie temporale V_{t-i} (dove *i* denota l'iesima effettuazione del loop), sia dei coefficienti C_i.

L'istruzione 1 del programma che controlla questo processore genera l'indirizzo di C_i. Ciò è spesso effettuato in un solo ciclo operativo, mediante la somma di una costante ad un valore di base. L'istruzione 2 genera l'indirizzo degli elementi V_{t-i}. Questi indirizzi sono poi usati per accedere ai dati e alle costanti, che vengono presentati all'ingresso del moltiplicatore durante le istruzioni 2 e 3. L'istruzione 4 effettua la moltiplicazione e porta il prodotto all'ingresso della unità aritmetica. Quest'ultima aggiunge il prodotto alla somma parziale, depositata in un registro temporaneo. Questi passi sono ripetuti in continuazione (loop) per calcolare X_t, il cui valore viene alla fine registrato nella memoria principale.

Come mostrato nella fig. 9, ogni unità del sistema è controllata da un proprio campo dell'istruzione; inoltre, per la maggior parte del tempo, esse operano contemporaneamente sui dati, anche se non gli stessi. Per esempio, V_{t-i} e C_i possono essere estratti dalle memorie simultaneamente, accrescendo significativamente l'efficienza del sistema. Le frecce mostrano come il flusso dei dati segua il concetto del pipelining e come questo sia legato al parallelismo. Pipelining e parallelismo migliorano certamente l'efficienza dell'elaboratore; tuttavia la capacità di un tipico elaboratore non va oltre i 5 o 6 milioni di moltiplicazioni al secondo. Questa è una velocità certo elevata, ma, come si è detto, essa è del tutto insufficiente in molte applicazioni "sonar".

Può sembrare ovvio che si possa ottenere un incremento di prestazioni collegando diversi elaboratori di segnali tra di loro; questa strada è stata però sinora assai poco battuta. Bill Wren, del Honeywell Marine Systems Center, per primo riconobbe che le caratteristiche dei calcoli richiesti per l'elaborazione di segnali potevano accordarsi con una particolare struttura di sistema a molti processori, capace di velocità molte volte superiori a quelle attualmente ottenibili. Il suo rappresenta un nuovo e assai promettente approccio al problema dell'elaborazione dei segnali.

4. L'approccio con bus a matrice

L'architettura di un calcolatore specificamente progettato per l'elaborazione dei segnali deve prendere le mosse dalle caratteristiche delle funzioni matematiche usate in tale campo di applicazioni. In un sistema di elaborazione di segnali "sonar", il flusso dei dati consiste di una molteplicità di canali (da 1 a 50), che attraversano successivamente parecchie unità funzionali (tipicamente da 6 a 12), ciascuna delle quali esegue sempre lo stesso tipo di operazioni matematiche su ciascuna serie temporale, indipendentemente dal loro valore. Pertanto, la struttura dell'elaborazione è statica per lunghi periodi di tempo. Ciò consente di realizzare opportuni compromessi nel disegno del sistema, come l'assegnazione di una data funzione elaborativa ad ogni specifico processore nonché semplificazioni nella gestione del sistema. Con ciò, il tempo di supervisione normalmente associato con sistemi multi-processore viene molto ridotto, consentendo l'interconnessione di un maggior numero di processori. Buona parte dell'onere di supervisione può, inoltre, essere trasferito nel periodo iniziale di operazione del sistema, minimizzandone l'impatto durante l'elaborazione vera e propria. Le rimanenti funzioni di controllo vengono poi suddivise tra i vari processori.

Un limite fondamentale al numero di processori del sistema può derivare dal meccanismo di interconnessione fisica tra i vari moduli. Un "bus" convenzionale non è in grado di trasmettere l'intenso scambio di dati tra i vari processori richiesto da questo tipo di elaborazione. Un bus convenzionale consente infatti la comunicazione simultanea solo tra due processori (fig. 10). Sarebbe come se gli operai che lavorano ad una linea di montaggio avessero a disposizione un unico carrello con cui trasportare i pezzi. Le differenti unità di elaborazione ("gli operai") debbono richiedere il bus ("il carrello"), avere il permesso di usarlo, usarlo ed infine restituirlo. La capacità del bus viene saturata rapidamente e le prestazioni del sistema non possono più aumentare, una volta raggiunto tale limite.

La natura delle elaborazioni sui segnali e la struttura a "pipeline" del sistema, rendono possibile assegnare specifiche funzioni a particolari processori e connettere i processori stessi mediante un "bus ad anello". In questo schema, una o più funzioni sono assegnate a ciascun processore ed ognuno di questi è collegato al bus mediante un registro, in cui si può leggere e scrivere. Il bus opera in sostanza come un registro circolare a scorrimento; in ogni singolo trasferi-

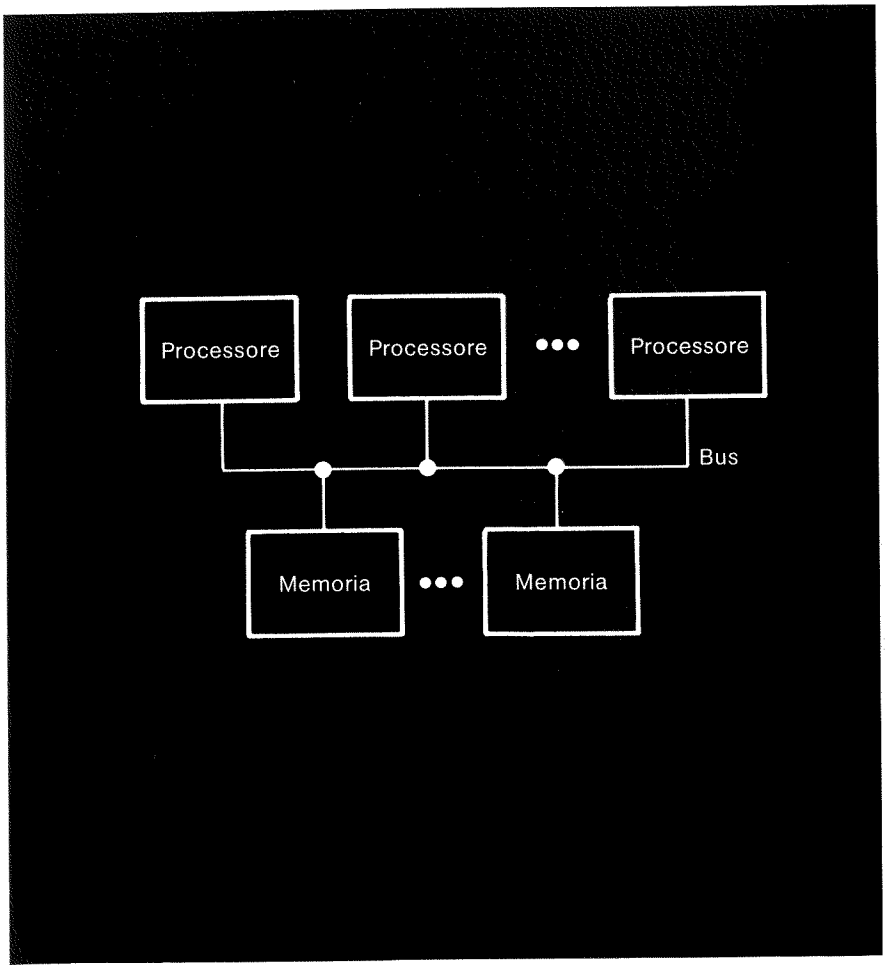


Fig. 10 - Un sistema multiprocessore con un "bus" convenzionale consente l'accesso alla memoria ad un processore per volta.

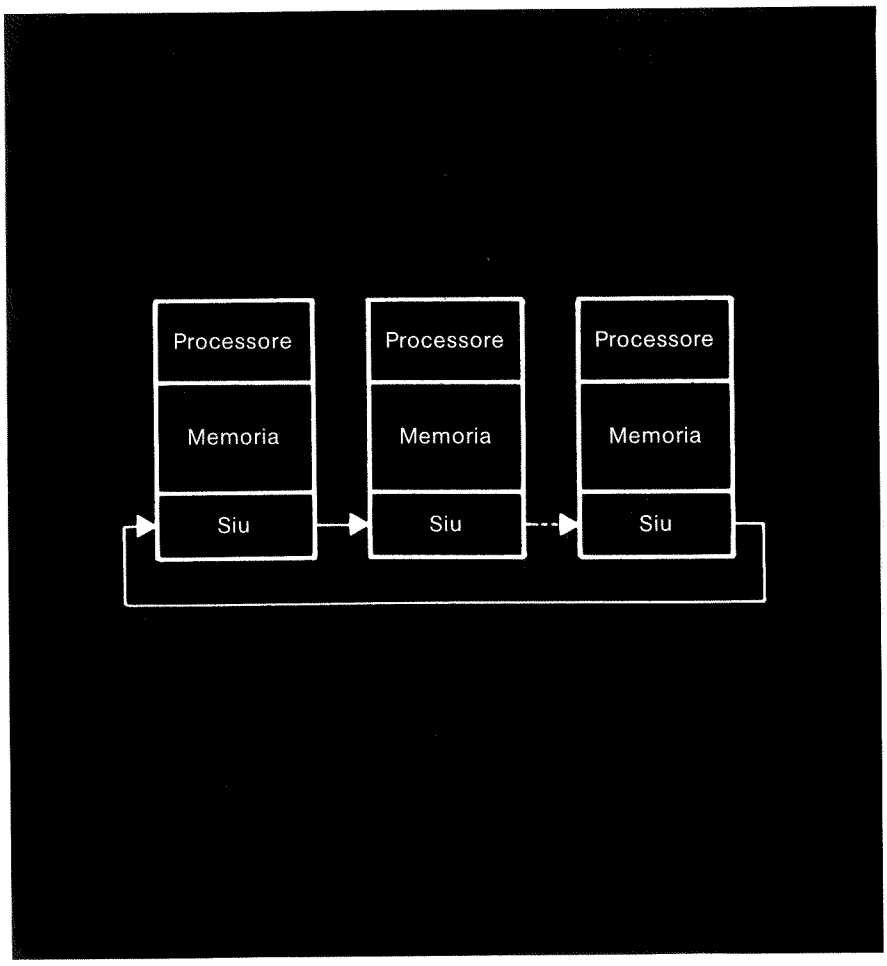


Fig. 11 - In un sistema multiprocessore con un bus "ad anello", ogni processore comunica col bus attraverso una unità di interfaccia (SIU: System Interface Unit). Un dato viaggia sul bus da un processore all'altro in modo analogo ad un prodotto che viaggia da una posizione all'altra lungo una linea di montaggio.

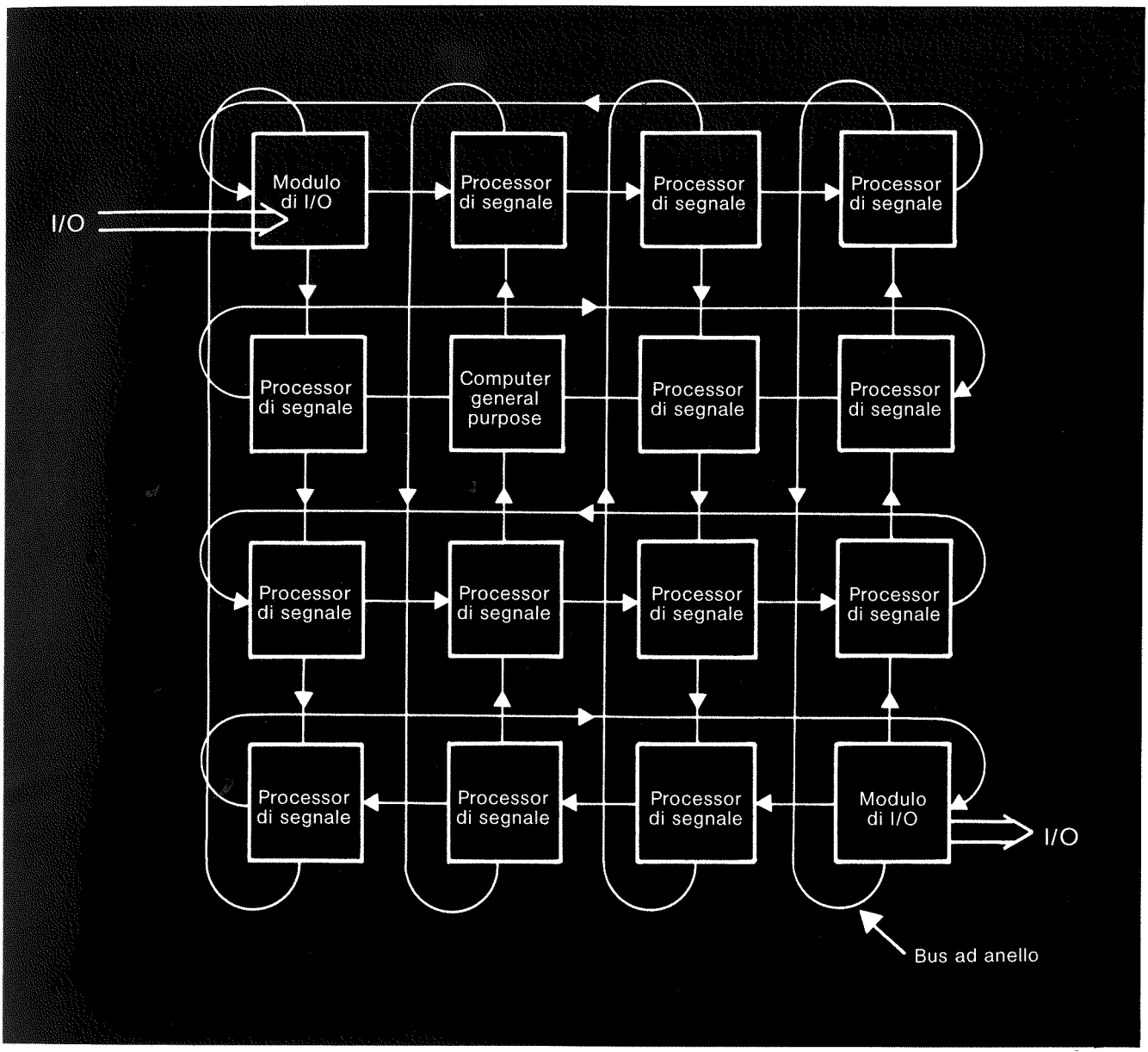


Fig. 12 - Un bus "a matrice" è formato dalla interconnessione di bus ad anello tra loro ortogonali e consente il trasferimento dei dati lungo linee perpendicolari.

mento, i dati non possono andare che da un processore al successivo, ma si possono avere tanti trasferimenti simultanei quanti sono i processori (fig. 11). Nel nostro paragone, ciò equivale a sostituire sulla linea di montaggio il carrello con un nastro trasportatore. Poiché nella elaborazione di segnali i dati sono generalmente inviati da un processore a quello adiacente, un bus ad anello aumenta sostanzialmente il numero di processori che possono essere interconnessi.

In una variante del bus ad anello, chiamata "bus a matrice", ciascun modulo di elaborazione è connesso a due bus ad anello tra loro ortogonali (fig. 12). Il bus a matrice è progettato in modo da consentire di deviare i dati da un percorso ad un altro; in tal modo, il tempo speso per trasferimenti non adiacenti viene ad essere sostanzialmente ridotto. In conseguenza un maggior numero di moduli può essere interconnesso senza degradare la capacità del sistema. Inoltre, se ai dati trasmessi sul bus vengono aggiunte delle informazioni di controllo (quale un codice ciclico ridondante, per verificare l'integrità dei dati), possono essere individuati i processori che eventualmente si guastassero. Oltre a ciò, il bus a matrice può consentire di far ruotare automaticamente i dati attorno ai moduli guasti e di riassegnare le funzioni di questi ultimi ad altri moduli. Ricerche in questa direzione vengono attualmente condotte.

Finora abbiamo discusso le architetture multi-processore da un punto di vista generale. Il calcolo deve confermare la praticità di tale soluzione prima di passare alla effettiva realizzazione. I progettisti

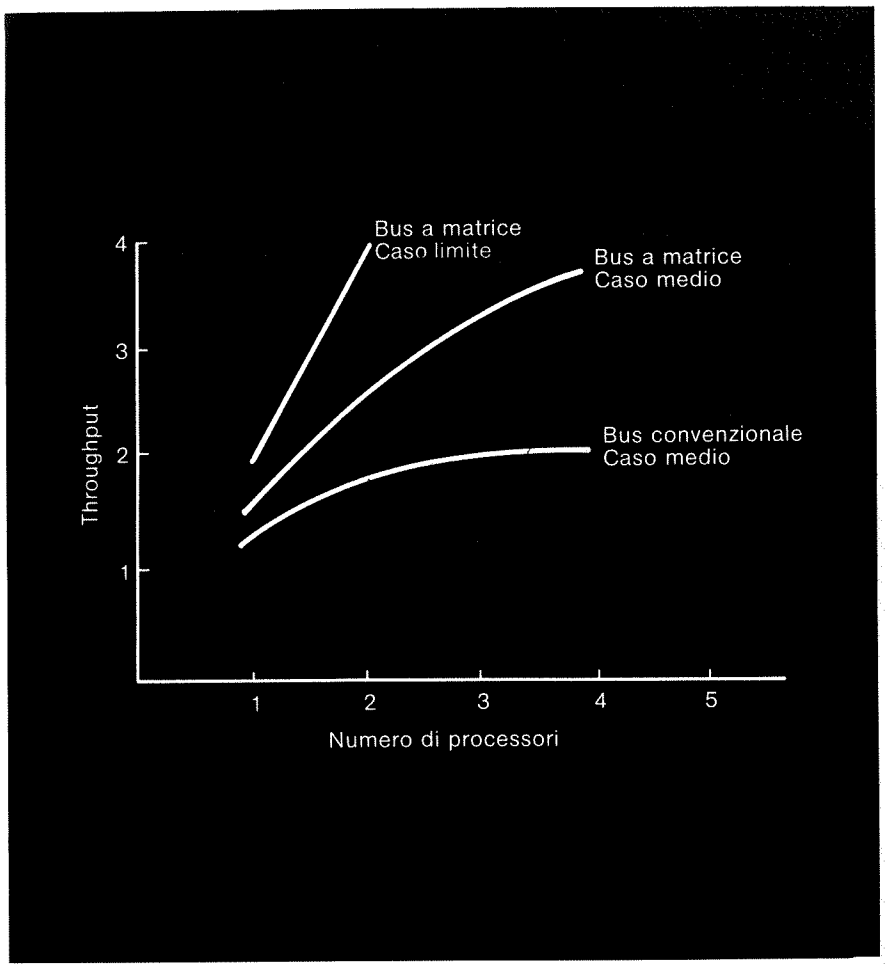


Fig. 13a) - Confronto tra le prestazioni di una architettura con bus a matrice e quelle di una architettura con bus convenzionale, nel caso si richiedano soprattutto attività di elaborazione.
(Caso limite = i trasferimenti avvengono tutti tra processori adiacenti).

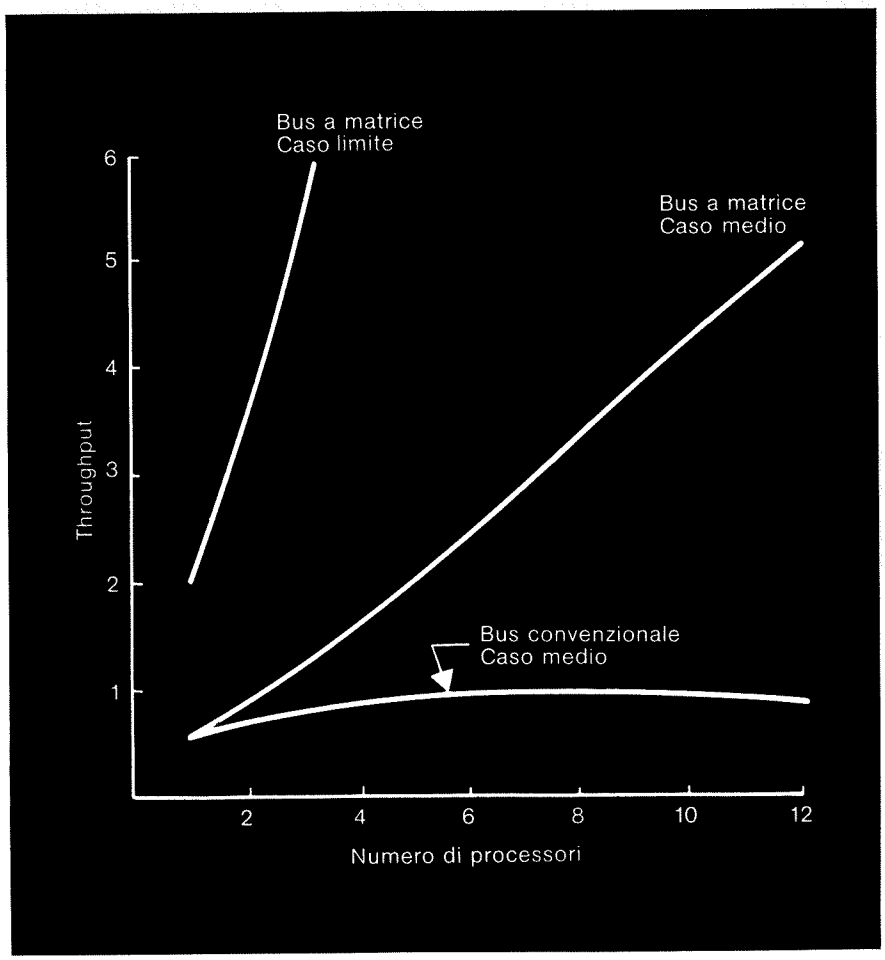


Fig. 13b) - I vantaggi di una architettura con bus a matrice diventano ulteriormente evidenti nel caso di intensa attività di ingresso/uscita.

di sistemi di elaborazione usano modelli analitici per stimare le prestazioni e quindi stabilire la praticità di nuove architetture. Nel caso specifico, si comincia con lo stimare il tempo ideale di esecuzione di una funzione, per passare poi a calcolare il tempo di supervisione ed i ritardi introdotti dalle operazioni di ingresso/uscita. Il tempo di esecuzione di una specifica funzione in un singolo processore è dato da:

$$E = (1 + \delta)\, E_i + I$$

dove E_i è il tempo di esecuzione ideale, assumendo cioè di avere l'uso esclusivo del processore e di non avere ingressi o uscite simultanei nè tempi di supervisione. Il termine δ rappresenta il tempo di supervisione mentre I è una funzione che misura il grado di sovrapposizione tra l'esecuzione della funzione e le operazioni di ingresso e uscita. Quest'ultime sono le operazioni di carico e scarico della memoria da parte dei dispositivi esterni, che degradano le prestazioni poiché sottraggono cicli di memoria; poiché questo può avvenire simultaneamente o indipendentemente dall'elaborazione, il termine I può essere zero oppure uguale al numero totale di operazioni di ingresso/uscita. Le definizioni dettagliate di δ e I escono dagli scopi di questo articolo. Determinato il tempo di esecuzione, l'inverso fornisce la capacità di elaborazione (throughput) del sistema:

$$T = \frac{1}{(1 + \delta)\, E_i + I}$$

Il modello di un sistema di elaborazione multi-processore è più complicato di quello con un singolo processore, cui si riferiscono le relazioni precedenti. I principali fattori di cui bisogna tener conto nei calcoli sono: il maggior tempo di supervisione che occorre spendere quando l'elaborazione è suddivisa in un numero di piccole funzioni, ciascuna delle quali può essere eseguita da un singolo processore (questo può significare che non viene sfruttata tutta la capacità del processore); il maggior tempo speso all'ingresso/uscita di ciascun processore, per comunicare con gli altri processori; e infine il ritardo generato dalla comunicazione attraverso il bus del sistema. Il modello precedente diviene quindi:

$$E = (1 + \delta + \gamma)\, E_i + I + I'$$

dove I' è un termine che tiene conto del ritardo causato dall'accesso al bus del sistema e γ della supervisione necessaria per distribuire le funzioni tra i vari processori. Mediante questo modello, è possibile dimostrare che una struttura con un bus a matrice ha una capacità di elaborazione assai più alta di quella di un sistema avente lo stesso numero di processori connessi mediante un bus convenzionale, anche quando sono richieste solo modeste attività di ingresso/uscita (fig. 13a). Quando è invece richiesta una rilevante attività di ingresso perché la funzione è spezzata in molte funzioni separate, la superiorità della struttura a matrice diventa ancor più evidente (fig. 13b). Se la struttura a matrice viene estesa in tre dimensioni, le prestazioni risultano ulteriormente esaltate.

5. Conclusioni

L'architettura qui presentata e discussa rappresenta un progetto di ricerca e sviluppo interno alla Honeywell, diretto a definire un sistema modulare di elevatissime prestazioni per l'elaborazione di segnali per impieghi futuri. Esso rappresenta un vero e proprio salto in avanti in questa area di sistemi. Questa architettura può rendere fattibile l'interconnessione di un numero elevato di moduli per ottenere altissime velocità operative. Attualmente non esistono in pratica sistemi multiprocessori o comunque elaboratori di segnali aventi una tale flessibilità. Il primo sistema multi-processore con bus a matrice, che costituisce una parte del progetto EMSP (Extendable Modular Signal Processor) del Marine Systems Center della Honeywell, sarà completamente definito entro il prossimo anno.

Connessioni con fibre ottiche nei sistemi di elaborazione

GIORGIO GRASSO
Soc. Cavi Pirelli
Milano
GIACOMO SCALZO
Honeywell Information Systems Italia
Centro di Ricerca e Progettazione
Pregnana (Milano)

1. Introduzione

È ormai universalmente riconosciuto che le fibre ottiche rappresentano uno dei più rivoluzionari progressi tecnologici nel campo della trasmissione dell'informazione.

Fin da quando le fibre a bassa attenuazione e i dispositivi elettroottici sono usciti dalla fase di sperimentazione, il mercato associato alla nuova tecnologia ha assunto cifre importanti e sta crescendo a ritmo sostenuto per raggiungere secondo una delle stime più autorevoli (Gnostic Concepts Inc.) la cifra di 1500 M$ alla fine degli anni '80, investendo non soltanto il campo delle telecomunicazioni, che attualmente trae i maggiori vantaggi dalla trasmissione ottica ma anche altri settori quali le applicazioni in campo militare, consumer, trasmissione dati e altri, in ognuno dei quali possono individuarsi esigenze attuali o emergenti che le caratteristiche delle fibre possono brillantemente coprire.

Analizziamo i principali vantaggi che la trasmissione ottica può vantare rispetto al tradizionale cavo in rame.

- Larghezza di banda molto grande: una sola fibra ottica può sostituire un numero elevato di portanti tradizionali.
- L'attenuazione molto bassa consente di realizzare lunghe tratte di collegamento senza rigenerare il segnale o, per collegamenti molto lunghi, di utilizzare un numero di ripetitori molto inferiore a quello necessario con tecnologia tradizionale.
- Peso e dimensioni molto contenuti permettono di realizzare cavi di ridotte dimensioni equivalenti per capacità trasmissiva a ingombranti e pesanti cavi metallici.
- La trasmissione ottica è immune da ogni disturbo di natura elettromagnetica e a sua volta non genera tale tipo di radiazione; le comunicazioni su fibra non presentano problemi di diafonia e risultano difficilmente intercettabili o disturbabili.
- I cavi ottici sono di tutta sicurezza in ambienti esplosivi o infiammabili, eliminando completamente il rischio di cortocircuiti e di scintille.

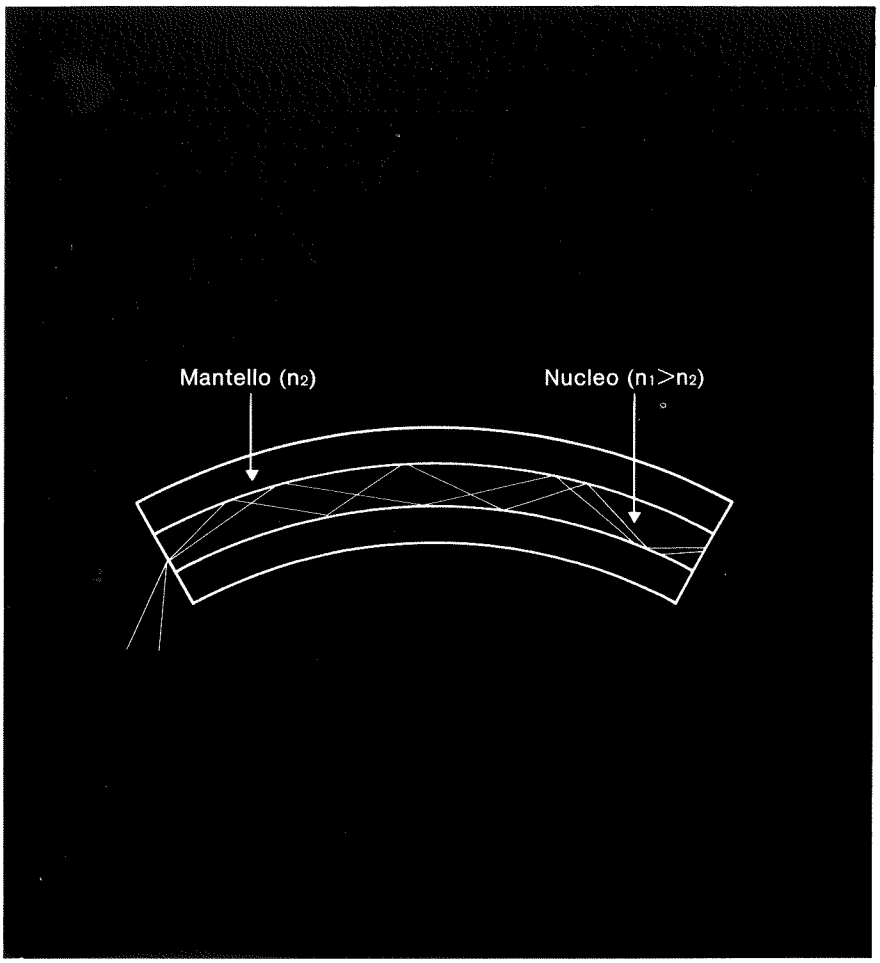
Vengono qui esaminati i principi alla base del funzionamento delle fibre, delle sorgenti di luce e dei rivelatori ottici e viene data una breve panoramica sulle principali tecniche di fabbricazione delle fibre e dei cavi.Segue un'analisi sulle probabilità che hanno le fibre ottiche di affermarsi nel campo della trasmissione dati in sistemi EDP; è questo il campo di applicazioni non relativo alle telecomunicazioni che offre le prospettive più interessanti come è confermato da tutte le proiezioni previsionali che anzi considerano il settore dello scambio dati tra computer come quello destinato ad assumere il secondo posto nel mercato complessivo e caratterizzato dal tasso di crescita più elevato nell'immediato futuro, quando cioè, l'emissione di una normativa meccanica ed elettrica, e ulteriori progressi sul fronte dell'affidabilità e dei costi renderanno convenienti e possibili tutte quelle applicazioni basate su produzioni in larga scala.

2. Le fibre ottiche

La fibra ottica è una guida d'onda dielettrica costituita da un cilindro di materiale vetroso (nucleo o "core") attorno al quale è un tubo concentrico e solidale (mantello o "cladding") costituito anch'esso in generale da materiale vetroso, con indice di rifrazione leggermente inferiore a quello del cilindro interno; nella sezione del core l'indice di rifrazione può essere costante (fibra "step index") o variabile in modo graduato (fibra "graded index"). Tale struttura implica che i raggi di luce incidenti sulla sezione di ingresso del core con angolo di incidenza sufficientemente piccolo vengano guidati all'interno del core stesso per effetto della riflessione totale all'interfaccia core/cladding (fig. 1).

In realtà, i possibili cammini ottici all'interno del core sono in numero limitato e corrispondono, nella

Fig. 1 - Propagazione di raggi luminosi in fibra ottica.



teoria della propagazione elettromagnetica ai modi di propagazione.

La trasmissione di informazioni in fibra ottica si attua modulando, generalmente a impulsi, una portante a frequenza ottica generata da un Laser o un LED. La presenza di diversi percorsi possibili per la portante ottica all'interno del core comporta una deformazione del segnale modulante dovuta ai diversi tempi di transito sui vari percorsi. Tale deformazione si manifesta in generale come una distorsione di ampiezza di tipo "filtro passa basso" sul canale tramissivo relativo al segnale modulante e provoca perciò una limitazione delle capacità trasmissive della fibra stessa.

Tale distorsione è molto meno sensibile nelle fibre "graded index" che nelle fibre "step index" in quanto la graduazione dell'indice di rifrazione all'interno del core tende ad equalizzare i tempi di transito della portante ottica sui vari percorsi. I migliori risultati si ottengono con un andamento dell'indice di rifrazione nel core quasi parabolico, che permette, per una determinata lunghezza d'onda, la completa equalizzazione dei tempi di transito.

Allo scopo di definire in modo completo questo nuovo mezzo trasmissivo sono state normalizzate una serie di grandezze legate sia alle sue proprietà fisiche sia alle sue caratteristiche trasmissive.

Le più importanti tra loro sono le caratteristiche dimensionali, la apertura numerica, l'attenuazione e la dispersione.

Caratteristiche dimensionali

La sezione della fibra ottica è definita geometricamente da due circonferenze corrispondenti l'una al core l'altra al cladding. I parametri più importanti che caratterizzano la geometria di una fibra sono perciò i diametri dei due cerchi e le loro relative eccentricità. Tali grandezze e le loro tolleranze hanno una importanza fondamentale nel determinare la possibilità di accoppiare le fibre tra loro senza incorrere in grosse perdite. Le dimensioni sono,per le fibre più in uso, dell'ordine di 50-100 micron per il core e 130-200 micron per il cladding. È in corso di completamento in sede internazionale un lavoro di normalizzazione per unificare tali parametri e le tolleranze relative.

Apertura numerica

L'apertura numerica della fibra corrisponde al seno del massimo angolo con cui un raggio luminoso può incidere sulla testa d'ingresso del core per essere guidato all'interno di esso. Tale grandezza determina l'efficacia di accoppiamento della fibra con le sorgenti luminose. Nelle fibre "graded index" in genere l'apertura numerica è intorno a 0,2 mentre nelle fibre "step index" può arrivare a 0,5 che corrisponde ad angoli di 12° e 30° rispettivamente.

Attenuazione

L'attenuazione intrinseca di una fibra ottica dipende essenzialmente dai fenomeni d'assorbimento e di scattering.

L'assorbimento è causato dalla presenza, anche in minima quantità, di impurità come lo ione OH associato alla presenza di acqua e ioni metallici quali C Fe.

Lo scattering dipende, oltre che dalla presenza di inclusioni all'interno del core, della interazione della radiazione elettromagnetica con la struttura amorfa dei materiali che costituiscono la fibra. A seconda dell'importanza relativa di questi due fenomeni, l'attenuazione di una fibra ha un diverso andamento in funzione della lunghezza d'onda di trasmissione. In effetti, mentre l'attenuazione causata da scattering tende a diminuire rapidamente all'aumentare della lunghezza d'onda ($\alpha \sim \frac{1}{\lambda^4}$), l'assorbimento più importante, cioè quello dovuto allo ione OH produce picchi di attenuazione localizzati alle lunghezze d'onda di risonanza della struttura molecolare. In pratica l'attenuazione in funzione della lunghezza d'onda ha un andamento del tipo rappresentato in fig. 2, nel quale in generale si possono individuare 3 zone di minimo relativo di attenuazione, dette "finestre", centrate rispettivamente su 850 nm, 1300 nm, 1500 nm in corrispondenza delle quali i valori estremi di attenuazione citati nella letteratura tecnica sono 2 dB/Km, 0,5 dB/Km, 0,2 dB/Km rispettivamente.

Dispersione

Il fenomeno della dispersione in fibra ottica multimodale è dovuta essenzialmente a due cause: la dispersione intermodale e la dispersione cromatica.

La dispersione intermodale già illustrata precedentemente produce una distorsione di ampiezza nel canale trasmissivo del tipo "filtro passa basso". La larghezza di banda utilizzabile per la trasmissione d'informazione dipende fortemente dal profilo dell'indice di rifrazione all'interno del core. Nelle fibre "graded index" di migliore qualità la larghezza di banda a −3 dB è dell'ordine di 1500 MHz·Km mentre nelle fibre "step index" è intorno a 30 MHz·Km.

La dispersione cromatica dipende dalla variazione della velocità della luce con la sua lunghezza d'onda. Poiché in generale il segnale da trasmettere viene inviato su di una portante luminosa non monocromatica, le diverse porzioni del segnale che si distribuiscono sulle varie lunghezze d'onda impiegano tempi diversi per transitare nella fibra. Anche tale fenomeno provoca una distorsione di ampiezza del tipo "filtro passa basso" nel canale trasmissivo.

L'entità di questa dispersione dipende dal tipo di materiale che costituisce la fibra, dalla lunghezza

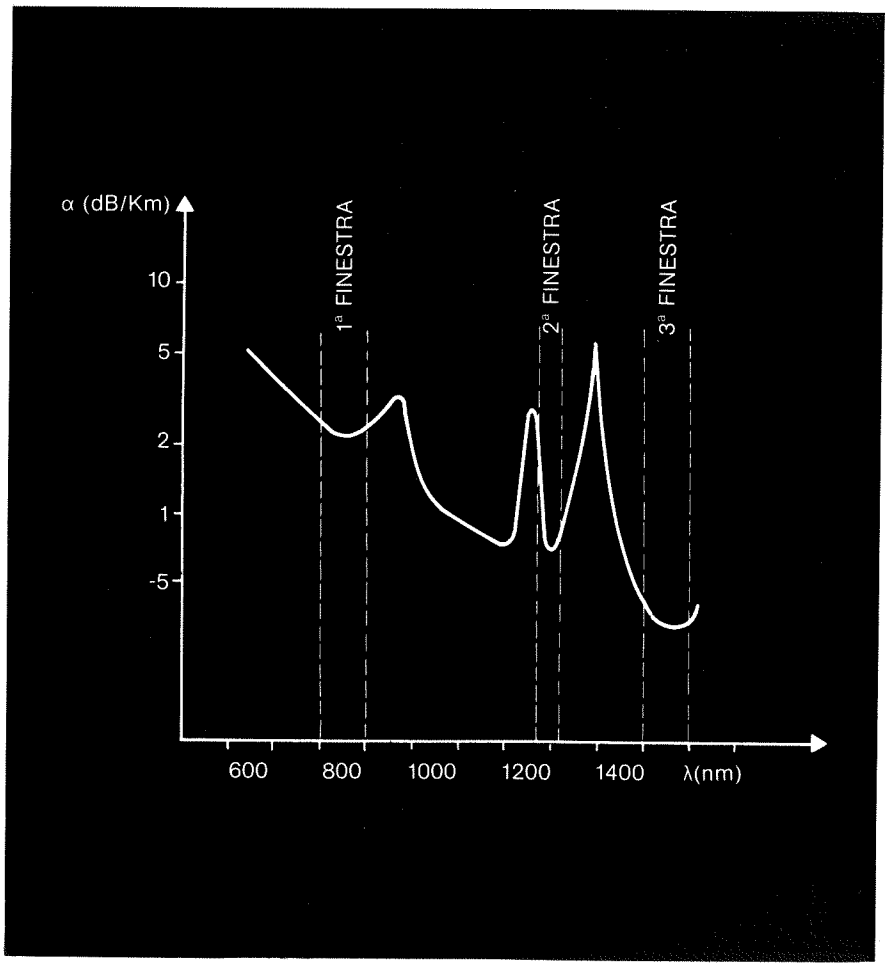


Fig. 2 - Andamento tipico dell'attenuazione di una fibra ottica in funzione della lunghezza d'onda della radiazione luminosa.

d'onda di picco e dalla larghezza dello spettro di emissione della sorgente luminosa utilizzata.

Ad esempio, se consideriamo una fibra "graded index" costituita da silice drogata con germanio e con larghezza di banda dovuta alla multimodalità maggiore di 700 MHz·Km, la larghezza di banda effettiva misurata con un LED di lunghezza d'onda di picco pari a 850 nm e larghezza spettrale a metà altezza di 50 nm sarà di circa 80 MHz·Km.

3. Processi produttivi delle fibre ottiche

Dal punto di vista dei materiali di cui sono costituite, le fibre possono sostanzialmente dividersi in due categorie:

- Fibre costituite di silice pura con l'aggiunta di piccole percentuali di drogante quali germanio, boro, fosforo per variarne l'indice di rifrazione nel core e nel cladding;
- Fibre costituite da due silicati diversi (comunemente detti vetri) uno per il core e uno per il cladding.

Vi sono inoltre delle particolari fibre costituite da un core di silice e un cladding plastico e inoltre delle fibre tutte di materiale plastico. Notevole è anche la varietà dei processi tecnologici utilizzati per la produzione di fibre ottiche a seconda del tipo di materiale e della qualità di fibra richiesta.

Il principio comune sul quale si basa la produzione delle fibre costituite di vetro o di silice è che una barra di vetro, scaldata fino a diventare plastica, se tirata riduce la sua sezione proporzionalmente alla velocità di trazione conservando la forma originale.

Molti dei processi produttivi consistono perciò nella realizzazione di una barra, detta preform, tale che l'indice di rifrazione nella sua sezione abbia proporzionalmente la variazione desiderata sulla sezione della fibra e nella sua successiva filatura fino ad ottenere la fibra delle dimensioni richieste. La fase più critica della lavorazione è senz'altro costituita dalla preparazione del preform. Considerando la necessità di garantire un'estrema purezza dei materiali ed un dosaggio accurato dei droganti, si sono in generale adottate delle tecnologie dall'industria dei semiconduttori. Un esempio è costituito dal metodo C.V.D. (Chemical Vapour Deposition) sviluppato per primo dalla Corning Glass negli Stati Uniti e che ha permesso di realizzare fibre di elevate prestazioni.

La fibra prodotta con il metodo C.V.D. è costituita da silice pura con l'aggiunta di piccole quantità di drogante (germanio, fosforo, fluoro, boro) il cui effetto è di far variare l'indice di rifrazione della silice stessa in modo controllato.

Tali componenti vengono ottenuti come risultato di una reazione chimica che avviene in una zona opportunamente riscaldata tra composti gassosi preventivamente purificati.

I componenti si depositano sotto forma di particelle solide ed è quindi necessario un successivo riscaldamento che dà luogo al consolidamento delle particelle e alla formazione di uno strato compatto di sostanza vetrosa della composizione desiderata.

Si ottiene in tal modo una barra (preform) dalla quale si fila la fibra delle dimensioni volute.

Vi sono diverse possibilità di realizzare il metodo descritto: un tipo particolare di C.V.D. è schematizzato nelle figure 3.

In questo caso (fig. 3a) i gas fluiscono all'interno di un tubo di silice pura riscaldato da una fiamma e combinandosi, per effetto dell'elevata temperatura, depositano sull'interno del tubo strati di silice drogata con germanio, elemento che aumenta l'indice di rifrazione della fibra.

Aumentando progressivamente la percentuale di germanio si ottengono degli strati con indice di rifrazione crescente.

Una volta ottenuto lo spessore sufficiente, il tubo viene fatto consolidare e collassare riscaldandolo a temperatura più elevata e ne risulta quindi un preform sulla cui sezione l'andamento di indice di rifrazione è di tipo graduato (figura 3b).

Il preform viene infine filato fino ad ottenere la fibra delle dimensioni volute (fig. 3c).

Grazie all'estrema purezza dei materiali e all'elevata controllabilità dell'indice di rifrazione, il metodo C.V.D. consente di ottenere fibre di elevatissime prestazioni: attenuazione a 820 nm <2 dB/Km, dispersione $<0{,}5$ ns/Km.

L'industrializzazione del metodo C.V.D. necessita tuttavia di sofisticate apparecchiature di controllo che ne sembrano giustificare attualmente l'impiego solo per fibre a cui si richiede un'elevata qualità.

4. Cavi ottici

Le fibre ottiche hanno una resistenza a trazione intrinseca elevata; ciò che rende fragile la fibra è la presenza all'interno del materiale di cui è costituita di microfratture, le quali innescano la rottura della fibra anche con limitate sollecitazioni.

Tali microfratture, anche se assenti nella fibra appena prodotta, tendono a crearsi rapidamente e successivamente a ingrandirsi per effetto di abrasioni meccaniche e dell'umidità. È perciò necessario effettuare in linea durante la produzione della fibra un rivestimento di materiale plastico (ad esempio kynar o gomma siliconica o resina epossidica).

Tali materiali danno alla fibra un'elevata resistenza meccanica anche in ambiente umido. Il carico di rottura di una fibra deve essere comunque considerato una grandezza statistica poiché esso dipende dalla distribuzione casuale delle disomogeneità nella fibra stessa. In media per una fibra di diametro 120 micron, con opportuno rivestimento il carico di rottura è intorno a 1 Kg, ma la dispersione dei valori può essere notevole.

Le fibre protette con il rivestimento primario non sono tuttavia an-

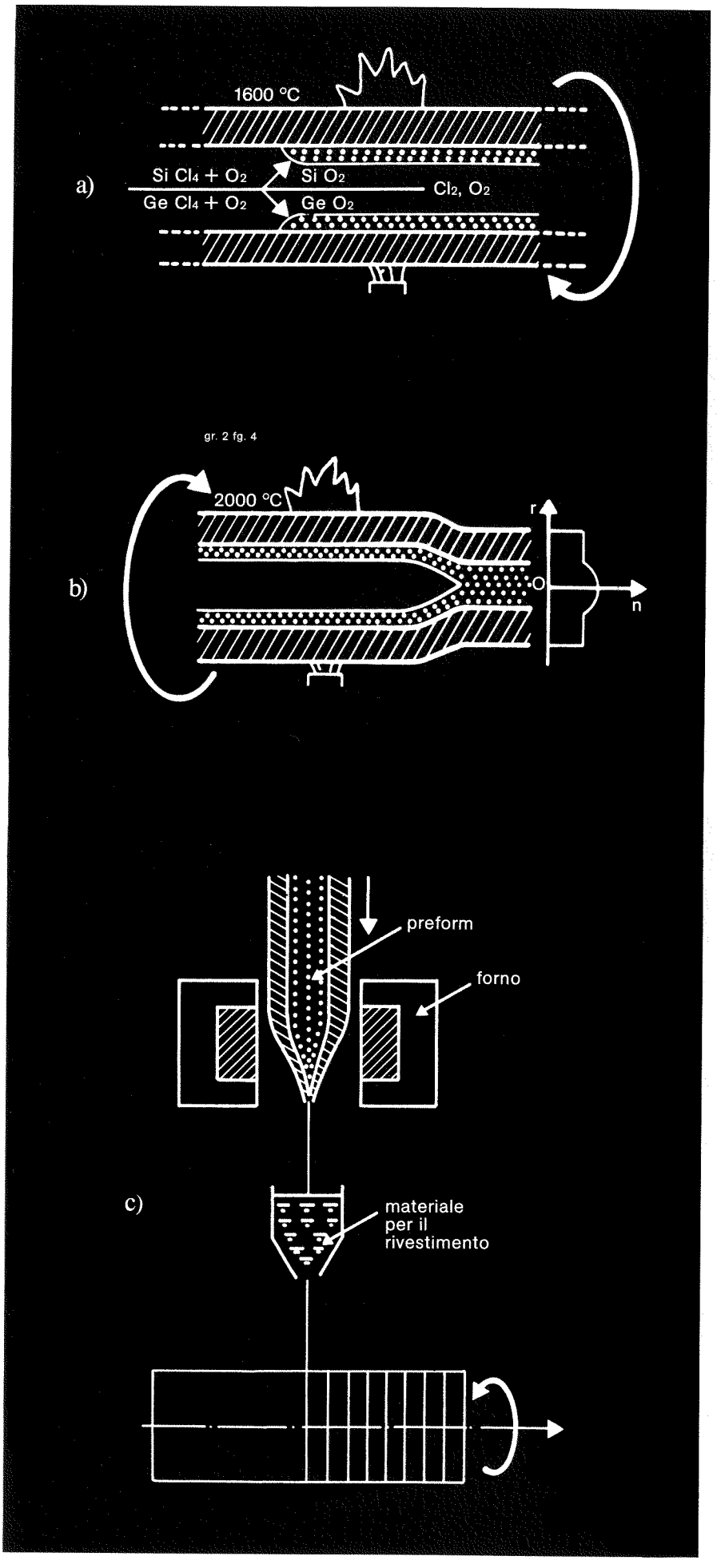


Fig. 3 - Fabbricazione di fibre ottiche col metodo CVD.
a) Deposizione di particelle di silice e biossido di germanio.
b) Consolidamento e collassamento del preform.
c) Filatura della fibra.

cora idonee ad essere utilizzate nella pratica.

In particolare la fibra è soggetta a rotture per effetto di curvature troppo accentuate ed inoltre non risulta, date le sue dimensioni, facilmente maneggevole. Vi è inoltre il problema relativo alle microcurvature che si generano in una fibra pressata da forze trasversali contro una superficie ruvida.

Tali microcurvature provocano un aumento di attenuazione in trasmissione che può essere anche notevole se le deformazioni sono distribuite su gran parte della lunghezza della fibra.

È necessario, quindi, dare alla fibra un'ulteriore protezione meccanica per permettere la sua pratica utilizzazione.

Tale prestazione è in generale realizzata inglobando le fibre in cavi la cui struttura dipende essenzialmente dall'applicazione a cui essi sono destinati. La protezione esercitata dal cavo sulle fibre deve essere infatti commisurata al tipo di sollecitazione a cui il cavo verrà sottoposto durante la posa e l'esercizio. Una particolarità dei cavi ottici è che nel progettarli è necessario tenere conto sia delle sollecitazioni meccaniche assiali e laterali che delle sollecitazioni termiche.

Le variazioni di temperatura nel cavo possono produrre sollecitazioni meccaniche sulle fibre a causa della forte differenza di coefficiente di dilatazione termica tra la silice che costituisce la maggior parte della fibra ottica e i materiali plastici che sono utilizzati in notevole quantità nei cavi (10^{-6}/°C contro 10^{-4}/°C). In generale perciò nei cavi ottici sono sempre inseriti elementi di rinforzo aventi il compito di controllare le sollecitazioni assiali sia di trazione sia di compressione.

Tondini di acciaio o di vetroresina sono gli elementi di rinforzo più comunemente usati nei cavi ottici.

Sono state costruiti in varie parti del mondo cavi ottici con strutture anche molto diverse tra loro; esse si possono in prima approssimazione suddividere in due categorie a secondo del modo con il quale è protetta la fibra. Vi sono in effetti cavi nei quali viene adottata una protezione definita lasca e cavi nei quali viene adottata una protezione definita aderente.

Cavi con protezione lasca.

In tale tipo di cavi la fibra viene isolata da eventuali sollecitazioni esterne, sia longitudinali che trasversali, per mezzo di una struttura di elevata resistenza meccanica nella quale essa si dispone in modo lasco. Un esempio di tale struttura è il rivestimento individuale della fibra con un tubetto di materiale plastico ad alto modulo elastico di diametro interno largamente superiore a quello della fibra (5÷10 volte) (fig. 4) nella quale la fibra si dispone in modo elicoidale così da risultare sufficientemente isolata dalle sollecitazioni esterne.

Il vantaggio principale dei cavi che adottano un rivestimento di tipo lasco è costituito dal fatto che le fibre non sono entro certi limiti influenzate dalle sollecitazioni esterne sia meccaniche che termiche. D'altra parte le dimensioni dei cavi sono considerevoli rispetto alla fibra nuda ed inoltre la scarsa protezione delle fibre alle estremità rende difficoltose le operazioni di giunzione e l'applicazione dei connettori.

La principale applicazione dei cavi con protezione lasca sembra essere nel campo delle telecomunicazioni nel quale si richiede una elevata qualità e una quasi assoluta stabilità delle caratteristiche trasmissive delle fibre.

Diverse tra le strutture di cavo finora realizzate sono quelle che adottano questo tipo di protezione della fibra.

La soluzione più comunemente scelta è quella del cavo tradizionale con struttura concentrica. In un tale tipo di cavo le fibre, protette singolarmente con un rivestimento di materiale plastico, sono cordate attorno ad un elemento centrale in generale metallico.

Il cavo è completato da fasciature e da una guaina esterna in materiale plastico. Un esempio di cavo a 8 fibre prodotto dalla Pirelli per applicazioni nel campo delle telecomunicazioni è rappresentato in fig. 5.

Cavi con protezione aderente

In tali tipi di cavo la fibra viene inglobata in una struttura di elevata resistenza meccanica. Questa aumenta notevolmente la resistenza assiale della fibra ma in generale non è in grado di proteggerla a sufficienza dalle sollecitazioni trasversali che provocano microcurvature delle fibre stesse. È perciò spesso associato al rivestimento di materiale resistente uno di materiale a basso modulo elastico, il quale tende a smorzare l'effetto sulle fibre delle sollecitazioni trasversali.

Un esempio di protezione aderente è il rivestimento individuale di una fibra con una doppia guaina, una, all'interno, di materiale morbido e una, all'esterno, di materiale ad elevato modulo (fig. 6).

In generale i cavi nei quali le fibre sono protette in modo aderente hanno il pregio di una struttura molto compatta; inoltre le fibre sono protette in modo adeguato alle estremità e ciò rende più agevoli le operazioni di giuntaggio e l'applicazione dei connettori. Tuttavia poiché le fibre sono solidali con la struttura del cavo esse risentono di tutte le deformazioni che le vengono imposte dall'esterno.
Di conseguenza le caratteristiche trasmissive e lo stato di tensione delle fibre possono venire influenzati da sollecitazioni esterne sia termiche che meccaniche.

Si può quindi considerare tale tipo di cavo per le applicazioni industriali nelle quali si utilizzano fibre di non elevata qualità e non si richiede una assoluta stabilità delle loro caratteristiche trasmissive.

Esistono numerosi esempi di strutture di cavo nelle quali si è adottato una protezione di tipo aderente.

Anche in questo caso la soluzione più comunemente scelta è quella del cavo tradizionale di tipo concentrico. In tale tipo di cavo le fibre, protette singolarmente con un rivestimento aderente, sono cordate attorno ad un elemento di

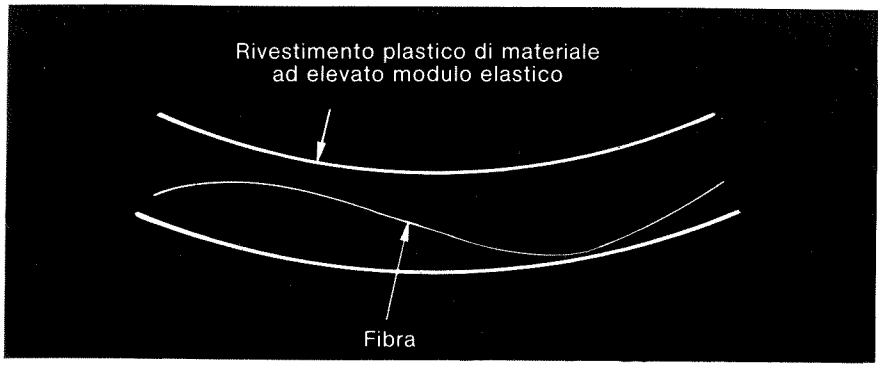


Fig. 4 - Fibra protetta singolarmente con rivestimento lasco.

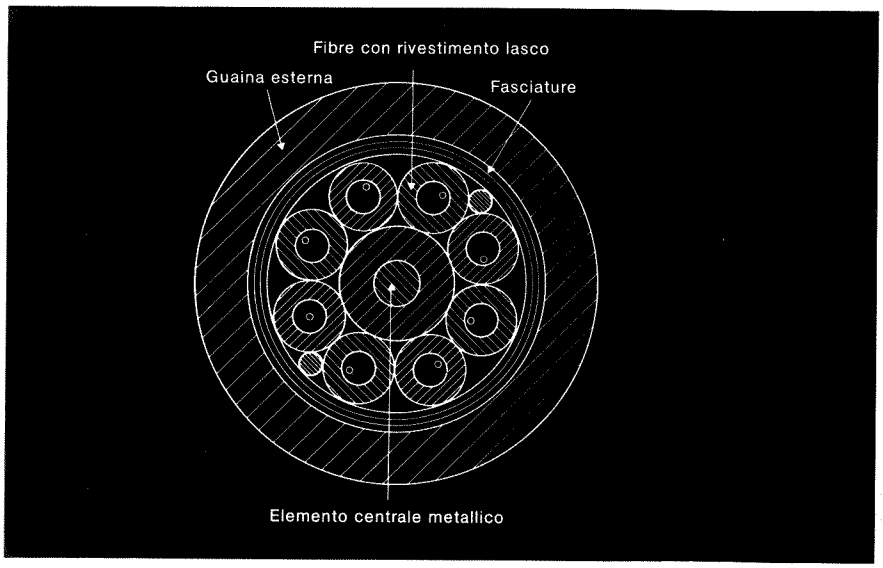


Fig. 5 - Sezione di un cavo a struttura concentrica con fibre rivestite singolarmente in modo lasco.

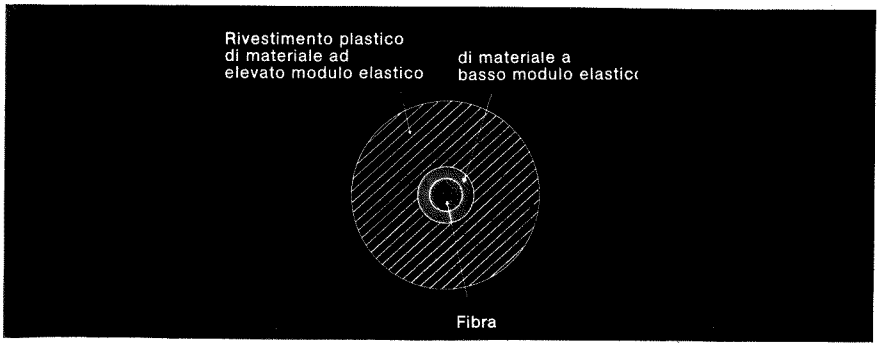


Fig. 6 - Fibra protetta singolarmente in modo aderente.

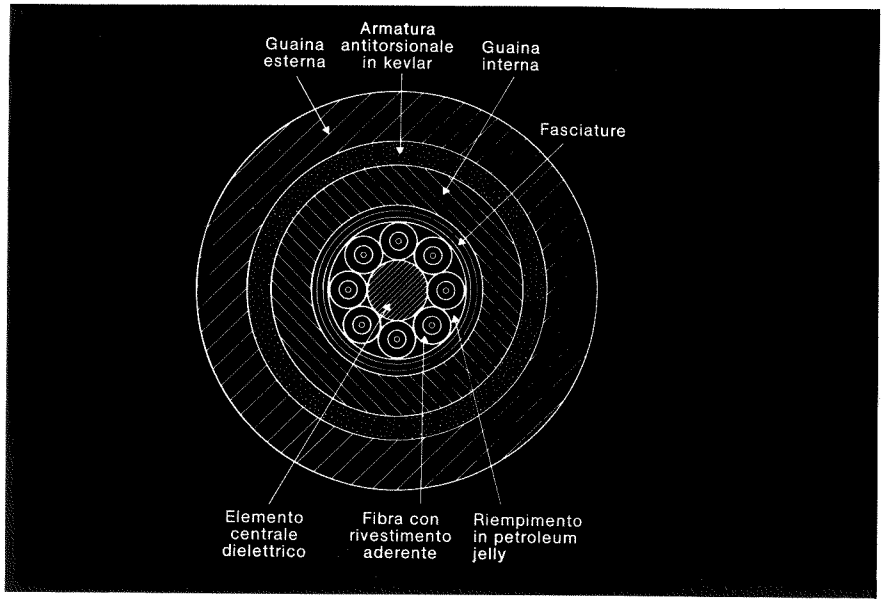


Fig. 7 - Sezione di cavo a struttura concentrica con fibre rivestite singolarmente in modo aderente.

rinforzo in generale non metallico come ad esempio un tondino di vetroresina. Il cavo è completato da guaine in materiale plastico nelle quali possono venire inseriti elementi resistenti tessili (Kevlar) o fili di vetro.

In fig. 7 è rappresentato un cavo a 8 fibre di questo tipo prodotto dalla Pirelli per trasmissioni via aerea.

Una soluzione molto particolare è costituita da cavi a nastro nei quali le fibre sono protette racchiudendole tra due nastri di materiale plastico che possono contenere un certo numero di fibre affiancate.

Più nastri possono essere associati tra loro con una leggera torsione longitudinale o possono essere avvolti attorno ad un elemento centrale resistente. Tale struttura si presta particolarmente alla realizzazione di cavi contenenti un numero elevato di fibre.

È stato ad esempio costruito un cavo del diametro di 12 mm contenente 144 fibre.

5. Sorgenti di luce

Tra le caratteristiche di una sorgente di luce, quelle più importanti per l'utilizzo in sistemi di comunicazione in fibra ottica sono:

la lunghezza d'onda della radiazione emessa; la potenza di emissione; la radianza, cioè la densità di potenza per unità di superficie e per angolo solido; la velocità di risposta; l'affidabilità.

Le uniche sorgenti di interesse per questi scopi sono i diodi ad emissione luminosa LED e i diodi laser a semiconduttore LD.

La fig. 8 mostra la ben nota struttura a bande di un semiconduttore. In un semiconduttore intrinseco, la maggior parte degli elettroni è confinata nella banda di valenza, tuttavia è possibile che elettroni acquistino energia sufficiente per saltare il "gap" di energia e portarsi nella banda di conduzione. Quando questo avviene vengono a generarsi due cariche libere, un elettrone e una lacuna, nella banda di conduzione e di valenza rispettivamente. Questi portatori di carica hanno tempo di vita limitata, vanno incontro cioè ad un processo di ricombinazione che le annulla.

Questo processo di ricombinazione è accompagnato da perdita di energia che può essere emessa sotto forma di fotoni e quindi di luce, o sotto forma di fononi (quanti di energia vibrazionale) e quindi in forma non radiativa.

La probabilità che si verifichi in maggior misura l'uno o l'altro fenomeno, è legato alle caratteristiche del materiale semiconduttore: il Germanio e il Silicio, dominatori incontrastati della tecnologia del transistore e dei circuiti integrati, sono poco utili per la realizzazione di dispositivi elettroluminescenti, mentre l'Arseniuro di Gallio e altri semiconduttori dei gruppi III e V possiedono tutte le caratteristiche necessarie per realizzare efficaci sorgenti di luce.

LED

I diodi ad emissione luminosa sono essenziamente delle giunzioni PN in cui elettroni e lacune sono iniettati nelle regioni P ed N rispettivamente, per mezzo dell'applicazione di una tensione di polarizzazione diretta (fig. 9).

Gli elettroni nella regione P e le lacune nella regione N rappresentano portatori minoritari e quindi con altissima probabilità di ricombinazione. Questo processo di ricombinazione, se il materiale base è stato scelto opportunamente, è associato ad emissione luminosa.

Un parametro che può essere indicativo per strutture elettroluminescenti di questo tipo è l'efficienza quantica interna, cioè il rapporto tra il numero di fotoni generati e quello dei portatori di carica iniettati. L'efficienza quantica per un LED varia tra il 20 e il 50% nei dispositivi tecnologicamente più sofisticati.

Importante è anche l'estensione della superficie emittente e il diagramma angolare di emissione, perché sia assicurato un buon accoppiamento ottico fra diodo e fibra.

Le condizioni per un buon accoppiamento, perché cioè la maggior quantità di energia luminosa venga spedita in fibra, sono legate, come è intuibile, al rapporto tra le dimensioni della superficie emittente e della sezione del core della fibra affacciata, nonché all'angolo del cono di divergenza della radiazione.

Sono state realizzate strutture in cui la superficie emittente è confinata in una piccola regione il cui diametro è di appena qualche decina di micron; tuttavia ridurre l'area di emissione non è sufficiente a garantire bassa perdita di accoppiamento se non si interviene opportunamente sull'angolo di divergenza della radiazione. La soluzione consiste nell'accoppiare al LED una minuscola lente convergente oppure posizionare la fibra tanto vicino alla superficie emittente da far si che essa raccolga la massima potenza emessa indipendentemente da quello che potrebbe essere il diagramma angolare di emissione del diodo impiegato.

Il posizionamento di una fibra ad una distanza dell'ordine del mm e anche meno dal diodo e perfettamente collimata, non può essere fatto dall'utilizzatore, ma è il costruttore stesso del diodo che accoppia al dispositivo in maniera permanente uno spezzone di fibra detto comunemente "pig-tail" (fig. 10) e al quale l'utilizzatore collegherà il cavo vero e proprio, rimandando così ad un collegamento fibra/fibra il delicato accoppiamento sorgente/fibra.

LASER

La luce prodotta da un LED è dovuta ad una emissione spontanea ed è quindi caratterizzata da grande incoerenza, i fotoni emessi, cioè, hanno valori di energia e direzione di propagazione differenti, per cui ne risulta uno spettro di emissione piuttosto ampio (tipicamente 40÷50 nm).

Quando i portatori minoritari attraverso la giunzione del diodo sono in numero molto elevato, può nascere il fenomeno dell'emissione stimolata che è alla base del funzionamento dei dispositivi LD. Quando un fotone colpisce un elettrone in uno stato eccitato può stimolare l'emissione di un altro

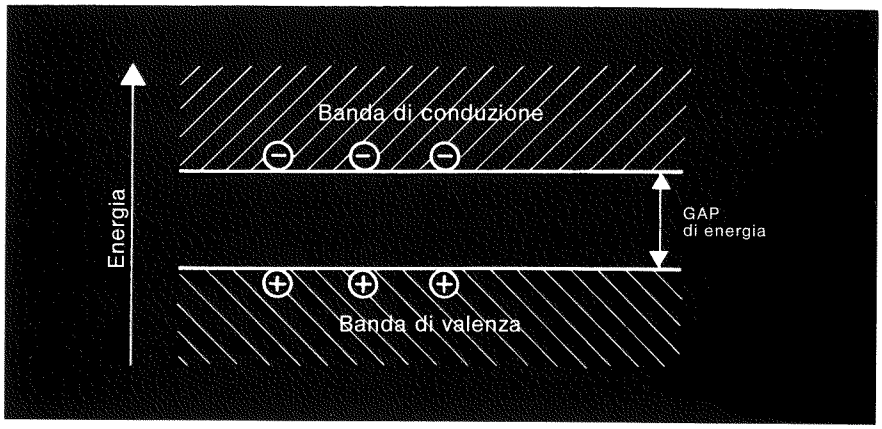


Fig. 8 - Struttura a bande di un semiconduttore.

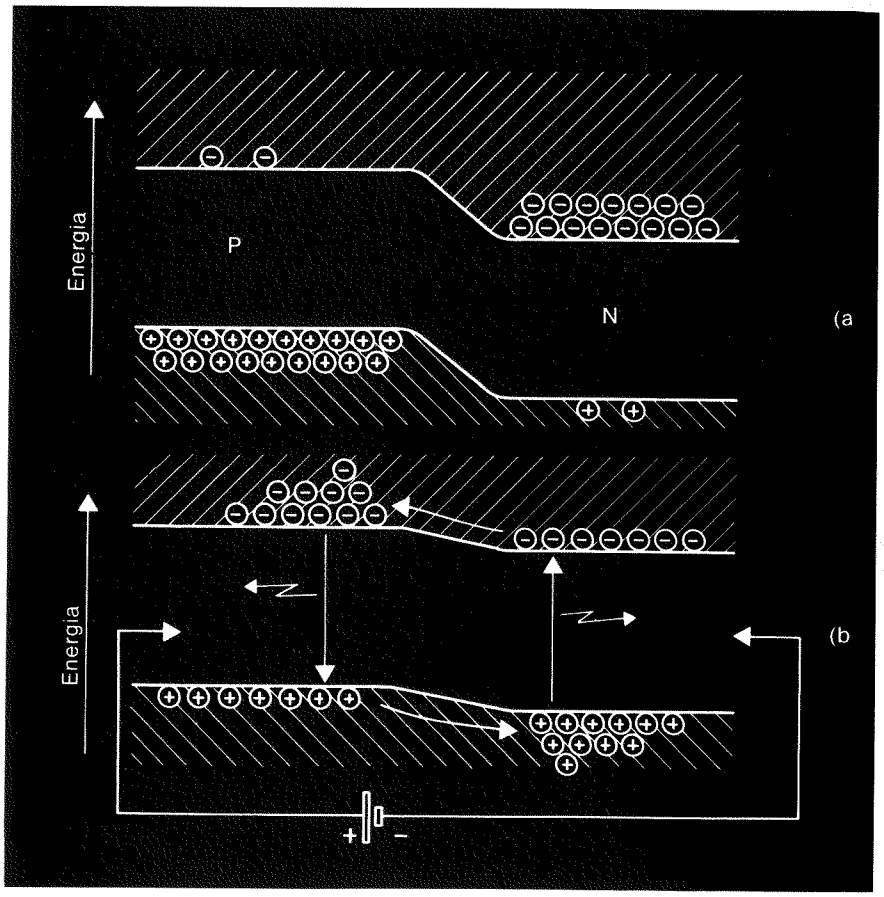


Fig. 9 - Diodo LED.
a) Il potenziale interno della giunzione rappresenta una considerevole barriera al moto degli elettroni e delle lacune.
b). L'applicazione della tensione esterna di polarizzazione diretta riduce notevolmente la barriera e consente l'iniezione di cariche.

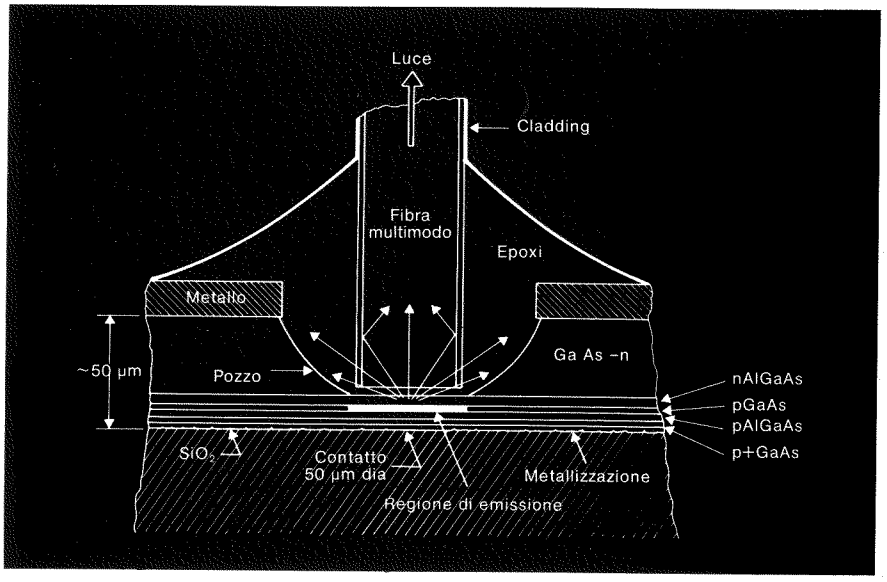


Fig. 10 - LED con "pig tail".

fotone caratterizzato da avere la stessa lunghezza d'onda, la stessa fase, e la stessa direzione del fotone stimolante; si innesca un fenomeno di amplificazione di luce. Per evitare il riassorbimento di questi fotoni dalle parti inattive del diodo, il dispositivo è realizzato in maniera da confinare la radiazione nella sola zona attiva con la creazione, mediante superfici semiriflettenti, di una vera e propria cavità ottica risonante.

Scaturiscono immediatamente i vantaggi che un diodo LASER a semiconduttore offre rispetto al più semplice LED. La radiazione elevatissima di un LD permette di immettere in fibra una potenza ottica notevole (1 mW e oltre), riducendo inoltre sensibilmente i problemi di accoppiamento con la fibra stessa anche del tipo a core molto piccolo (monomodale); la larghezza spettrale è molto contenuta (qualche nm) e riduce notevolmente la dispersione della fibra; la velocità di risposta, per il meccanismo fisico che genera l'emissione (stimolazione e non ricombinazione spontanea) è molto alta e permette frequenze di modulazione molto alte dell'ordine del Gb/s.

Per contro il LASER è un dispositivo la cui costruzione richiede processi molto critici e fortemente controllati; questo implica costi molto alti e una affidabilità decisamente minore di quella di un LED.

Inoltre, la caratteristica potenza/corrente altamente non lineare impedisce l'uso del dispositivo in sistemi analogici, e la dipendenza della temperatura impone soluzioni circuitali e di impiego molto complesse.

Il diodo LED è quindi preferito in tutti quei casi in cui il prodotto banda·distanza rientra nei valori che il LED stesso può soddisfare, e l'uso del LASER è limitato a quelle applicazioni dove requisiti di distanza e ampiezza di banda giustificano l'uso del dispositivo.

6. Rivelatori

Compito dei rilevatori è quello di riconvertire il segnale luminoso in arrivo dalla fibra in corrente elettrica di intensità proporzionale alla potenza ottica ricevuta.

Tra i vari dispositivi che potrebbero essere utilizzati, gli unici che per caratteristiche, dimensioni e praticità d'uso sono impiegati in sistemi di comunicazione in fibra ottica sono: i fotodiodi PIN e i fotodiodi ad effetto valanga APD.

Il principio alla base del funzionamento di questi dispositivi è l'effetto fotoelettrico su una giunzione PN polarizzata inversamente.

Se un raggio di luce di opportuna lunghezza d'onda colpisce un atomo di materiale N o P, il fotone assorbito può trasferire energia sufficiente ad un elettrone della banda di valenza da fargli compiere il salto del "gap" creando così due portori di carica di segno opposto, un elettrone libero e una lacuna.

La carica così creata, che sarà minoritaria, se riuscirà a raggiungere lo strato di svuotamento della giunzione senza ricombinarsi, (cioè se il fenomeno è avvenuto ad una distanza dello strato di svuotamento non superiore alla lunghezza di diffusione della carica stessa), attraverserà la giunzione e determinerà passaggio di corrente. Se l'atomo illuminato appartiene proprio allo strato di svuotamento, allora la probabilità che le cariche liberate contribuiscano alla corrente nel diodo è altissima, e nello stesso tempo anche il tempo di risposta del dispositivo è proporzionalmente più piccolo. È evidente a questo punto l'esigenza di avere diodi nei quali la zona di carica spaziale sia grande, in modo che possa raccogliere la massima parte dell'energia luminosa in arrivo senza creare grossi problemi di allineamento con la fibra.

Perché la zona di svuotamento sia grande occorre che le regioni P ed N del diodo siano poco drogate, e i dispositivi PIN sono appunto realizzati affacciando ad una zona P uno strato pochissimo drogato, praticamente intrinseco, al quale segue una zona di tipo N dove è realizzato il contatto ohmico. Il diodo PIN (così chiamato dalla situazione di drogaggio dei tre strati che praticamente lo compongono), è un dispositivo di facile impiego, non richiede alte tensioni di polarizzazione, ed è sufficientemente veloce (il tempo di risposta è dell'ordine di qualche ns). Un aspetto negativo è la bassa *responsività*, cioè l'attitudine del diodo a convertire potenza ottica in segnale elettrico; in un diodo PIN non supera il valore di 1 A/W e le variazioni di corrente in uscita devono essere opportunamente amplificate.

Se il sistema richiede valori di responsività maggiori, cioè se è necessario rivelare correttamente potenze ottiche molto basse senza far ricorso a forti amplificazioni, si deve far ricorso a fotodiodi di tipo ADP.

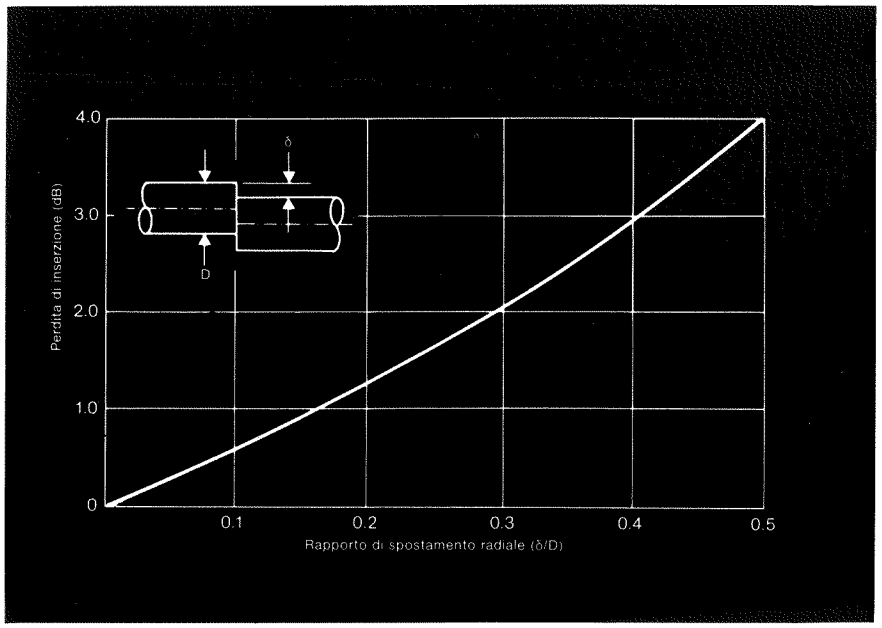
La caratteristica di questo tipo di rivelatore è quella di possedere un meccanismo interno di amplificazione di corrente. È composto da una giunzione PN alla quale è applicata una forte polarizzazione inversa.

Gli elettroni generati per impatto fotonico sono accelerati dal forte campo elettrico e possono acquisire tanta energia cinetica da riuscire ad eccitare altri elettroni in stati liberi e a moltiplicare con un processo a valanga il numero di portatori che contribuiscono alla corrente di giunzione.

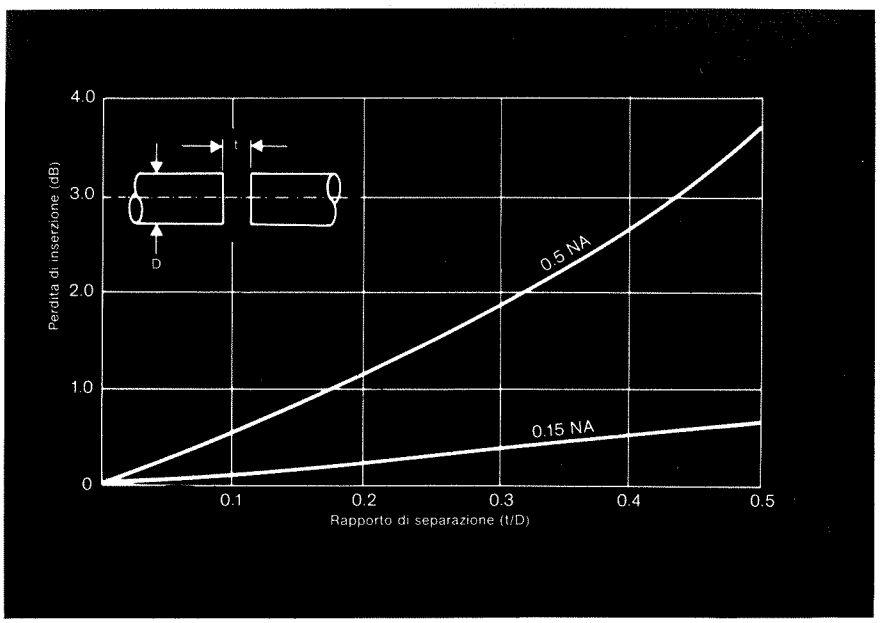
Lo svantaggio principale di un rivelatore ADP è la necessità di disporre dell'alta tensione necessaria alla forte polarizzazione inversa (150÷350 V).

L'uso di ADP in apparecchiature digitali standard richiede quindi l'utilizzo di convertitori di tensione DC/DC; inoltre il guadagno interno dipende fortemente dalla temperatura e questo implica ulteriori complicazioni circuitali.

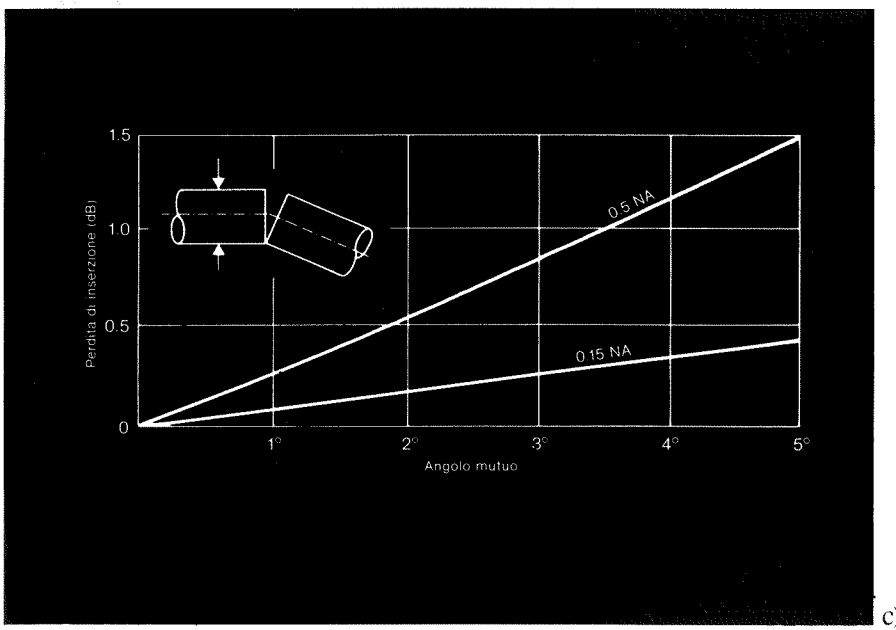
Il campo delle applicazioni degli APD è quindi limitato al campo delle telecomunicazioni dove occorre realizzare lunghi tratti di collegamento senza ripetitori e ad alta velocità di modulazione. Per tutte le altre applicazioni, il diodo PIN è indubbiamente da preferire.



a)



b)



c)

Fig. 11 - Connessione di due fibre ottiche. a) Effetto dello spostamento laterale. b) Effetto della separazione. c) Effetto del disallineamento angolare.

7. Giunti e connettori

Quando si affronta il problema della interconnessione fra due fibre emergono tutti gli svantaggi che le fibre ottiche possiedono rispetto al tradizionale cavo in rame.

Se infatti stabilire un collegamento elettrico tra due cavi in rame non presenta problemi particolari e può essere realizzato anche in condizioni di emergenza e con le tecniche più disparate, la natura stessa del segnale che viaggia in fibra richiede l'uso di attrezzature tecniche particolari.

L'obbiettivo è quello di ottenere una perdita di potenza ottica minima. Problemi costruttivi e di posa in opera, limitano a circa un chilometro la pezzatura massima per le fibre ottiche, mentre il settore delle telecomunicazioni richiede lunghi tratti che ormai si misurano in decine di chilometri.

Se l'attenuazione introdotta da una giunzione non è tenuta al minimo, nel bilancio complessivo delle perdite del collegamento possono rappresentare una quota molto alta e sproporzionata all'attenuazione dell'intera lunghezza di fibra che i più recenti progressi hanno portato al di sotto del dB/Km.

Possiamo dividere i sistemi di interconnessione in due categorie: giunti e connettori, intendendo per giunti quei sistemi che realizzano collegamenti permanenti e definitivi, e per connettori tutti i dispositivi, utilizzati soprattutto per connettere il cavo con le apparecchiature, che permettono un agevole e veloce collegamento e scollegamento. Nella realizzazione di questi dispositivi, i problemi sono essenzialmente di natura meccanica; collegare due fibre significa affacciarne le estremità libere o più esattamente la superficie dei due "core", e quanto sia un'operazione critica lo si può dedurre dall'analisi della fig. 11 in cui sono evidenziate le perdite che intervengono per cattivo posizionamento delle due fibre.

È evidente che basse perdite si possono ottenere solo realizzando dispositivi caratterizzati da tolleranze meccaniche estremamente contenute.

La fig. 12 mostra alcune tecniche realizzative per i giunti fissi; le possibilità sono due, o si procede ad una vera e propria saldatura delle estremità, o si collegano con l'ausilio di particolari meccanici che, variamente conformati, non sono altro che delle precise guide che assicurano l'allineamento delle fibre da connettere. Una goccia di collante, con indice di rifrazione opportuno, minimizza le perdite dovute allo strato d'aria tra le due sezioni e sigilla la giunzione proteggendola dalle impurità. Se ben realizzato, uno di questi giunti può introdurre un'attenuazione di solo qualche decimo di dB.

Più complessa è la realizzazione di connettori per connessioni "mobi-

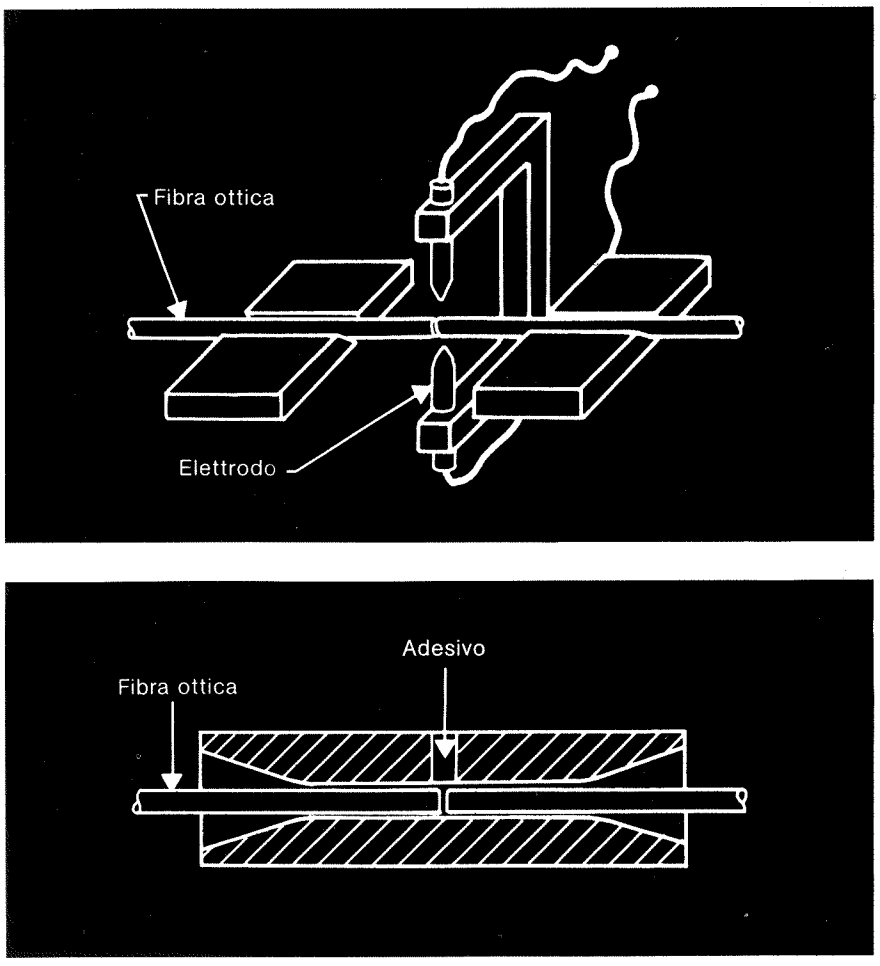


Fig. 12 - Realizzazione di giunti fissi. a) Giunzione per fusione con arco elettrico. b) Giunto a guida di centraggio.

li". Alle esigenze di bassa attenuazione si aggiungono quelle antitetiche, di mobilità, semplicità di montaggio, semplicità di connessione, ridotta manutenzione.

Anche in questo campo il mercato offre varie soluzioni.
Quando le prestazioni richieste sono basse, nel caso di collegamenti a distanza limitata e a bassa velocità, si può far ricorso a connettori del tipo più semplice, in qualche caso realizzati in plastica, e facilmente montabili sulla fibra che non richiede particolari preparazioni, e comunque tutta l'operazione può essere svolta dal diretto utilizzatore con l'ausilio di un minimo di attrezzatura.

I connettori destinati a sistemi con prestazioni superiori sono caratterizzati da soluzioni costruttive più accurate, (utilizzo di rubini per centrare le fibre, ecc.), materiali molto stabili dal punto di vista deformazioni termiche, e, naturalmente, da un costo molto maggiore. Maggiore cura è anche posta nella preparazione delle estremità terminali delle fibre da connettere; la sezione deve essere rigorosamente a 90° rispetto all'asse e lappata e pulita con speciali attrezzi.

8. Prospettive di applicazione delle fibre ottiche nel campo dei sistemi di elaborazione dati

A differenza dei collegamenti nel campo delle telecomunicazioni in cui,per i grandi vantaggi offerti dalle fibre ottiche, è facile prevedere per il futuro un uso generalizzato della nuova tecnologia, per i collegamenti nell'ambito di un sistema di elaborazione il vasto campo di prestazioni e le caratteristiche richieste ai cavi impongono un'analisi accurata, che tenga conto degli aspetti economici e implementativi specifici.

La tabella di fig. 13 riassume schematicamente la situazione per le quattro classi di applicazioni in cui si può schematicamente suddividere l'insieme dei collegamenti nel settore EDP.

Di tali interconnessioni sono evidenziate le lunghezze tipiche, il modo di trasmissione (parallelo o seriale) e il data rate, cioè la quantità di informazioni trasferita nell'unità di tempo con le tecnologie tradizionali. Questi tre parametri non sono indipendenti tra di loro.

Nel caso di trasmissione seriale, in cui cioè il dato è trasferito inviando su di un unico filo in modo sequenziale tutti i bit che lo compongono, il data rate dipende dalla cadenza con cui questi bit vengono presentati sulla linea, e questa cadenza è fortemente condizionata dalle caratteristiche elettriche del cavo, in ultima analisi dalla lunghezza.

Fig.13 - Tecnologie di interconnessione nei sistemi di elaborazione dei dati.

Classi di connessioni	Tecnologia attuale	Modo	Data rate	Lunghezza	Caratteristiche	Tendenza
Connessioni interne	Wire - Wrap Piste Stampate Flat Cable	P P P		Qualche metro	Limitazioni derivano, in particolare, da fenomeni di diafonia.	La tendenza dei circuiti ad essere sempre più veloci esalta i problemi indicati. Nell'ambito delle tecnologie tradizionali è necessario ricorrere a configurazioni particolari, complesse e costose.
Connessioni locali tra CPU e terminale o periferica lenta	Cavo Coppia twist Cavo coax	S S S	19.200 b/s 9.600 b/s 1 M b/s	15m 300m 2 Km	Forte dipendenza della velocità dalla distanza. Per basse velocità di trasmissione e brevi distanze le tecnologie tradizionali sono molto vantaggiose economicamente.	Le periferiche e i terminali richiedono sempre maggiori velocità di trasferimento dati e a distanze decisamente superiori a quelle oggi permesse.
Connessioni tra CPU e periferiche veloci	Coppie twist in configurazione Flat Cable	P	Può arrivare a velocità dell'ordine del centinaio di Mb/s	15 m	L'alta velocità di trasferimento dati richiesta si realizza ricorrendo alla trasmissione in modo parallelo su molti fili.	Svincolo della periferica a risiedere vicino alla CPU. Riduzione del volume dei cavi di collegamento richiesta dalla miniaturizzazione del sistema.
Connessioni computer-computer per reti locali Connessioni computer-terminale remoto	Modem	S	Qualche centinaio di Kb/s		La necessità del Modem anche per distanze relativamente brevi (qualche Km) limita notevolmente le prestazioni.	La forte spinta verso una informatica distribuita impone l'introduzione di tecniche che permettano velocità di scambio dati maggiori.

Quando le prestazioni richieste sono decisamente elevate è indispensabile ricorrere alla trasmissione parallela, in cui cioè più linee operano in parallelo per trasferire simultaneamente altrettanti bit, per esempio un'intera parola o un dato ancora più articolato.

Anche in questo caso, comunque, le distanze che possono essere coperte sono molto limitate, anche per problemi legati alla maggiore complessità del cavo che si traduce in problemi di installazione e costo più alto.

La tabella riassume la situazione attuale e mette in evidenza le problematiche che possono emergere in un futuro molto prossimo; in questo contesto il ricorso alle caratteristiche offerte dalle fibre ottiche appare sicuramente interessante, ma non sempre la sostituzione sarebbe realmente vantaggiosa.

Nel campo dei collegamenti interni al computer vero e proprio, in cui i trasferimenti dei dati sono caratterizzati da alta velocità, perché legata direttamente alle prestazioni dell'intero sistema, sono i problemi di diafonia che potrebbero spingere verso l'impiego di segnali ottici; ma le esigenze di velocità ne escludono l'impiego. Le tecniche usate attualmente, poiché operano ad alto parallelismo e su lunghezze di linea molto contenute, garantiscono una velocità di operazione estremamente elevata, e comunque superiore a quella che potrebbe essere raggiunta con collegamento ottico, nel bilancio del quale occorre inserire oltre al tempo di propagazione della fibra (estremamente contenuto), anche il ritardo introdotto dai dispositivi elettroottici di conversione e dal relativo hardware.

Non bisogna dimenticare inoltre i problemi meramente tecnici legati ad una realizzazione di questo tipo, soprattutto per quanto riguarda la distribuzione del segnale ottico tra i vari elementi che compongono il sistema; allo stato attuale dell'arte, tale distribuzione può essere effettuata solo utilizzando componenti piuttosto complessi e costosi.

La situazione è diversa se si considerano le linee esterne al computer vero e proprio, cioè i collegamenti rivolti alle periferiche e ai terminali, o le linee di interconnessione tra i nodi di una rete di elaborazione. In queste applicazioni sono maggiormente sentite le limitazioni proprie dei cavi metallici. La limitata larghezza di banda e la forte dipendenza di questa dalla lunghezza del cavo, impone l'uso di linee parallele su molte vie quando la quantità di dati da trasferire nell'unità di tempo è elevata (periferiche veloci), oppure impone velocità di trasmissione basse quando la distanza da coprire è elevata (terminali remoti).

In questo contesto le fibre ottiche offrono grandi vantaggi e, dato il valore del prodotto larghezza dibanda x lunghezza, possono far prevedere possibilità molto interessanti.

Nel collegamento delle periferiche veloci è possibile eliminare gli ingombranti cavi piatti con sottili cavi ottici, eliminando nello stesso tempo le pesanti limitazioni in distanza che oggi impongono l'istallazione delle periferiche molto vicine all'elaboratore. Inoltre, i cavi ottici aprono il campo, finora vietato, dei trasferimenti di dati ad alta velocità su lunghe distanze, permettendo di sfruttare a pieno le possibilità elaborative di terminali intelligenti o di incrementare l'efficienza delle reti di elaborazione.

Al di là dei vantaggi enunciati, tuttavia, il costo dei sistemi ottici, e la scarsa maturità della tecnologia con tutte le implicazioni che comporta (impiego difficile, necessità di personale specializzato, mancanza di standard, ecc.) rendono l'introduzione della trasmissione ottica nel campo dei sistemi di elaborazione piuttosto lenta.

Un altro fattore importante che contribuisce sensibilmente a rallentare l'affermazione delle fibre ottiche in campo EDP è la scarsa flessibilità che hanno i protocolli di dialogo utilizzati attualmente per adattarsi ad una trasmissione di tipo seriale quale necessariamente deve essere quella ottica.

Questi protocolli sono stati sviluppati senza prevedere la possibilità di una serializzazione, per cui la conversione può essere realizzata solo con pesanti modifiche alle apparecchiature, in molti casi tali da inficiare alcuni dei vantaggi delle fibre ottiche e introdurre un costo aggiuntivo proibitivo.

In tempi brevi, i motivi che possono spingere all'uso della nuova tecnologia vanno ricercati in quelle che sono le caratteristiche uniche delle fibre, caratteristiche cioè che derivano dalla natura fisica del mezzo e che quindi non possono riscontrarsi nei cavi in rame. In questo senso, di grande interesse è in particolare l'immunità alle interferenze elettromagnetiche offerta dai cavi ottici.

Con il diffondersi dei sistemi di elaborazione, aumentano sempre più le applicazioni che impongono alle linee di collegamento condizioni ambientali particolarmente sfavorevoli; con le fibre ottiche si riescono ad ottenere valori di BER (Bit Error Rate) tipici di $10^{-12} \div 10^{-13}$ anche nell'ambiente più rumoroso, dove l'utilizzo di cavi di rame porterebbe seri problemi di schermatura.

Un'altra caratteristica esclusiva delle fibre ottiche che sta assumendo sempre maggior importanza è rappresentata dalla impossibilità di captare il segnale che viaggia in linea senza interrompere il collegamento o quanto meno senza alterarlo profondamente. Questa caratteristica è particolarmente utile quando è necessario "proteggere" l'informazione, necessità sempre più sentita con l'estendersi dei settori applicativi del computer in campo bancario, amministrativo, diplomatico, militare, ecc.

Da quanto precede, si può dedurre che esistono le premesse per un uso significativo delle fibre ottiche nel settore EDP, anche perché determinate esigenze, oggi ancora limitate, sono destinate ad accentuarsi.

Il costante progresso tecnologico dovrebbe realizzare nei prossimi anni le condizioni tecniche ed economiche necessarie perché le fibre ottiche trovino effettiva applicazione nei sistemi di elaborazione dei dati.

Informatica e psicologia (*)

BERNARD MELTZER
Progetto Intelligenza Artificiale
Politecnico di Milano

Mi sembra che ci sia una logica quasi inevitabile nello sviluppo della scienza dei calcolatori che porta direttamente al cuore della psicologia. Questo fenomeno presenta un certo numero di aspetti, alcuni dei quali verranno qui presentati e discussi.

1. Lo spettro dei programmi

Cominciamo con un problema di programmazione. Consideriamo la scrittura di un programma per invertire una matrice numerica. Ne può risultare un programma ingegnoso ed efficiente se il programmatore ha usato opportunamente le sue cognizioni di algebra lineare e di calcolo numerico.

Tuttavia, questa base di cognizioni è nella sua testa, non nel programma; pertanto, pur essendo molto efficiente, un programma di questo tipo è rigidamente specialistico e privo di flessibilità, cioè sa fare una cosa e soltanto quella.

Però le esigenze sia della società industriale moderna sia della creatività dei suoi scienziati ed ingegneri non sono soddisfatte, ora e tantomeno in futuro, da sistemi aventi generalità e scopi così limitati. Così, per tornare all'esempio della matrice, potremmo invece preparare un programma che incorpori alcune delle conoscenze di algebra e di calcolo numerico ai quali mi sono riferito e le usi per eseguire l'inversione delle matrici.

Permettetemi di definire un simile programma come più "cognitivo" di quello precedente.

Naturalmente, anche in tale caso il programmatore dovrà usare ulteriori conoscenze, e cioè un linguaggio di programmazione e modi efficienti per effettuare la codifica. Si potrebbe immaginare un programma ancor più "cognitivo", che incorpori queste ulteriori conoscenze e sia in grado di effettuare da solo l'inversione di matrici.

Quanto maggiore è la conoscenza incorporata in un programma (di cui la maggior parte sarà naturalmente di tipo generale), tanto più esso sarà versatile.
Ad esempio, uno dei programmi sempre più cognitivi immaginabili potrebbe non solo invertire una matrice ma anche risolvere equazioni polinomiali con coefficienti numerici.
Questa è una serie immaginaria di programmi; non so se qualcuno abbia realmente scritto un programma che usi le sue conoscenze generali di algebra lineare per invertire matrici, sebbene esista almeno un programma nel campo della matematica (di cui parlerò più tardi), che fa cose anche più notevoli. Ma questo illustra il concetto che esiste uno spettro di possibili programmi, più o meno cognitivi, in tutti i campi della programmazione. Quelli al livello più alto sono generalmente classificati come intelligenza artificiale, sebbene, come credo abbia fatto notare Nils Nilsson, il confine tra questi ed i programmi "ordinari" si sposta costantemente verso l'alto man mano che la scienza degli elaboratori progredisce. In ogni caso, non è una distinzione molto netta perché anche un programma molto "poco cognitivo", come quello elementare dell'inversione di matrice da cui siamo partiti, non è più privo di "intelligenza" del noto metodo per calcolare la radice quadrata di un numero che ci è stato insegnato a scuola, e che soltanto una piccola parte di noi sarebbe in grado di giustificare pienamente.

Cio che succede quando un programmatore mette sempre più conoscenza nel suo programma (occorre ricordare che la conoscenza non è solo di fatti ma anche di procedure) è che egli trasferisce una parte sempre maggiore del suo pensiero nel programma, in una forma di rappresentazione. In altri termini, il materiale della psicologia cognitiva viene inserito nei programmi con una rappresentazione più o meno esplicita. È vero che sono incluse principalmente

(*) Il materiale di questo articolo è stato tratto da un intervento dell'Autore dal titolo „Is computational science relevant to psycology?", presentato al Convegno Annuale AICA 1980.

solo quelle parti di materiale che sono aperte all'introspezione; ma la comprensione delle parti "inconscie" rimane uno dei problemi principali della psicologia nel suo complesso e più avanti in questa esposizione presenterò alcune considerazioni sul contributo che la scienza degli elaboratori potrebbe dare alla soluzione di tale problema.

2. Rilevanza e irrilevanza

Molti studiosi contemporanei di psicologia, filosofia e altre materie sostengono che i programmi degli elaboratori non possono rappresentare, o presentare, fenomeni mentali simili a quello umano, per tutta una serie di ragioni. Vorrei discuterne alcune.

Una delle obiezioni più comuni è che la mente umana funziona sulla base di substrati biologici costituiti da materiale neuronico, ben diverso dai transistori, microcircuiti, film magnetici, ecc. e che perciò il modo in cui si realizza la funzione cognitiva deve essere del tutto diverso da quello dei programmi degli elaboratori. È probabile che solo poche delle persone che si occupano di elaboratori siano molto colpite da queste argomentazioni, perché per loro è scontato che uno stesso programma che gira su due macchine diverse, aventi dispositivi e architettura interamente differenti, esegue sostanzialmente nello stesso modo il compito assegnato.

Già molto tempo prima che fosse inventato l'elaboratore elettronico, il più grande psicologo del nostro secolo, Sigmund Freud, aveva sostenuto l'irrilevanza di questa obiezione. In un suo scritto [1], ad un lettore immaginario che gli chiede che cosa intenda per "apparato mentale" e da che cosa sia costituito, risponde nel modo seguente: "Sarà presto chiaro che cos'è l'apparato mentale; ma io vi devo pregare di non chiedere di quale materiale sia fatto. Non è questo un argomento di interesse psicologico. Alla psicologia ciò è altrettanto indifferente quanto, ad esempio, per l'ottica il fatto che le pareti di un telescopio siano di metallo piuttosto che di cartone".

Sebbene Freud abbia ragione, questa non è comunque l'ultima parola sulle relazioni tra psicologia e neurofisiologia, e in seguito vorrei dire qualcosa in merito.

Un'altra obiezione comune viene fatta ad un livello che stà sopra l'hardware. Ci si riferisce al fatto che i programmi "cognitivi" dell'intelligenza artificiale svolgono compiti intellettuali essenzialmente attraverso la ricerca esaustiva, e sono pertanto profondamente diversi dal metodo umano che usa scorciatoie, intuizioni, immaginazione, "ispirazione", e così via. Questo punto di vista è così lontano dalla verità che vorrei ora dare alcuni esempi interessanti, che lo confutano in modo spettacolare.

3. Programma di backgammon di Berliner

Come tutti sanno, esistono molti programmi di giochi che vengono eseguiti dall'elaboratore contro giocatori umani con diversi livelli di successo. Per esempio, il programma di scacchi 4,5 di Slate e Atkin ha vinto il campionato open del Minnesota, negli USA. I programmi di scacchi di maggior successo si basano sostanzialmente sull'esame dettagliato di possibili sequenze di mosse a vari livelli di profondità, ma non sono quella analisi "cieca" ed esaustiva immaginata da alcuni critici; infatti si può facilmente dimostrare che una simile analisi sarebbe del tutto inefficace in termini pratici. Ciò che viene fatto è il tipo di analisi delle possibilità eseguito normalmente da giocatori umani, siano o no campioni; in questo ambito i programmi di scacchi potranno spesso essere in grado di operare in modo più completo del giocatore umano. Ma forse è più interessante il caso del backgammon (in italiano "tric-trac").

Nel luglio 1979 a Montecarlo il campione mondiale di backgammon, l'italiano Luigi Villa, fu battuto per 7 a 1 in un incontro con il programma di Hans Berliner BKG 9.8 (Backgammon 9.8). Berliner, che è un ricercatore sull'intelligenza artificiale alla Carnegie-Mellon University di Pittsburg, stese il programma [2, 3] sulla base di una brillante estensione del principio usato in uno dei primi programmi di giochi, quello di Samuel per la dama. In questo approccio si usa ben poca ricerca sistematica; le mosse sono basate invece su una valutazione generale, mediante un numero finito di elementi caratteristici del gioco. Così nella dama questi elementi possono comprendere la possibilità di mangiare una pedina dell'avversario o di fare dama, mentre nel backgammon (un gioco di cui non sono pratico) si potrebbero includere fattori come la capacità di difesa dei pezzi, il loro grado di mobilità, la vulnerabilità dell'avversario.

Il programma di Samuel per il gioco della dama aveva raggiunto un buon livello già molti anni fa, e quando citavo la brillante estensione che Berliner aveva fatto di questo programma, mi riferivo al maggior realismo e alla accuratezza del suo modo di rappresentare la valutazione (quello che in termini umani chiameremo il "giudizio") della situazione sulla scacchiera.

Dopo aver analizzato le sequenze dell'incontro con Villa, Berliner trasse le seguenti conclusioni: "Non c'è dubbio che BKG 9.8 abbia giocato bene. Villa ha fatto un gioco tecnicamente corretto per quasi tutta la partita, mentre il programma non ha fatto il gioco migliore in 8 su 73 situazioni. Soltanto uno di questi errori ha messo il programma in difficoltà. BKG 9.8 ha però fatto faville quando si è trattato di usare l'immaginazione. Uno che non sapesse chi fosse il Nero e il Bianco, avrebbe potuto pensare che era l'uomo a fare il gioco brillante e la macchina quello piatto".

4. Sfogliando i giornali

Il linguaggio è stato definito la finestra della mente, e molti programmi sono stati scritti, usando differenti approcci, per compren-

dere, parafrasare, tradurre e riassumere testi di linguaggio naturale. Per dare un'idea di ciò che è stato fatto finora, prenderò come esempio un programma che setaccia notizie di giornale, cercando cose che l'interessino e dando un breve riassunto dei fatti in una lingua qualsiasi, come inglese, russo o spagnolo.

Il suo nome è FRUMP (Fast Reading and Understanding Memory Program) e fu sviluppato dal gruppo per la comprensione del linguaggio naturale di Roger Schank all'Università di Yale [5, 6]. L'approccio seguito, detto di "dipendenza concettuale", si basa sulla rappresentazione del contenuto del testo di partenza in termini di concettualizzazioni correlate, piuttosto che sulla analisi grammaticale, che giuoca un ruolo relativamente minore in questo programma. Che FRUMP non trovi più difficile riassumere una notizia di cronaca inglese in russo o in cinese piuttosto che in inglese, lo si deve al fatto che la rappresentazione adottata è interlinguistica.

Ecco un esempio delle capacità di FRUMP, applicato alla seguente notizia presa da un giornale americano:

"Un violento terremoto ha colpito l'Italia Settentrionale ieri sera, facendo crollare interi quartieri in città a NE di Venezia, vicino al confine Yugoslavo, uccidendo almeno 95 persone e ferendone almeno 1000, secondo quanto ha riferito il Ministero degli Interni italiano. Un portavoce governativo ha detto che, nella sola città di Udine, si teme vi siano almeno 200 morti sotto le macerie. La città, sulla linea ferroviaria principale tra Roma e Vienna, ha una popolazione di circa 90.000 abitanti. Il portavoce dei carabinieri, la polizia nazionale paramilitare, ha detto che si hanno notizie di gravi danni da mezza dozzina di città ai piedi delle Alpi, con intere famiglie sepolte dal crollo di edifici. Risultano ancora interrotte le comunicazioni tra molti centri della zona. Il terremoto è stato di 6.3 gradi della scala Richter, che misura i movimenti tellurici. In aree abitate, una scossa di 4° grado può provocare danni moderati, una di 6° grado può essere grave, una di 7° indica un terremoto con effetti disastrosi".

Ecco il riassunto fatto da FRUMP:

"95 persone uccise e 1000 ferite in un terremoto che ha colpito l'Italia. La scossa registrata è di 6.3 gradi della scala Richter".

Il riassunto in spagnolo dice:

"Hubo 95 muertos y 1000 heridos en un terremoto fuerte en Italia. El terremoto midió 6.3 en la scala Richter".

5. Creazione di concetti e scoperte in matematica

Mentre i programmi del gruppo di Schank sono centrati su concetti della vita di tutti i giorni, quello che segue si occupa invece di concetti matematici e, inoltre, genera esso stesso altri concetti e fa su di essi congetture più o meno interessanti.

Il programma è stato scritto dal giovane americano Douglas Lenat e descritto nella sua tesi presentata nel 1976 alla Stanford University [7, 8]. Precedentemente, molti studi erano stati fatti su programmi per la dimostrazione di teoremi matematici, ma in generale il loro paradigma era costituito da una base di dati costituita da assiomi e dall'applicazione di regole di inferenza nell'ambito di determinati sistemi logici. Ma io credo, e i lettori potrebbero controllarlo analizzando i loro processi mentali nella ricerca di soluzioni, ad esempio, di problemi di geometria elementare, che il più delle volte l'inferenza gioca un ruolo relativamente secondario nel pensiero matematico. Quest'ultimo è un'attività mentale molto più ricca: può implicare il procedere per intuizione, il cercare analogie, il fare esperimenti concettuali o esplorazioni empiriche con esempi e controesempi, usando e generando concetti che sono compresi solo parzialmente, facendo generalizzazioni, ed altro ancora. Non c'è quindi da sorprendersi se programmi legati troppo strettamente al paradigma di inferenza logica abbiano raggiunto rapidamente un tetto di capacità non particolarmente elevato.

Lenat si è concentrato sulla ricerca di modelli relativi alla scoperta di concetti e proposizioni matematiche interessanti. In effetti il suo programma aveva ben poche capacità deduttive, nel senso stretto di dimostrazione di teoremi. L'attività del programma consisteva nel tentare di eseguire una sequenza di compiti secondo il loro "grado di interesse", abbandonando un compito non appena le risorse di elaborazione assegnate divenivano insufficienti.

Sebbene la maggior parte delle scoperte fatte fosse relativa ai numeri naturali, il programma AM non aveva su di essi conoscenze di base (il nome AM sta probabilmente per "Matematico-Automatico"). Il programma le scoprì per conto suo, anche se mancò di scoprire concetti come i numeri reali e le frazioni.

La sua base di conoscenze di partenza consisteva in circa 100 concetti e circa 250 regole euristiche di quale azione intraprendere secondo le circostanze. I concetti includevano oggetti come insieme, lista (il programma era scritto in linguaggio LISP, orientato alle liste), tavole di verità, relazioni come appartenenza ed uguaglianza, operazioni come inversione, intersezione e composizione. Lenat sosteneva, basandosi su riferimenti al lavoro degli psicologi Piaget e Copeland, che la sua raccolta di concetti di partenza intendeva essere quella di un bambino di 4 anni, ma non so quanto ciò sia vero.

Le regole euristiche erano rappresentate da ciò che in gergo si chiamano "produzioni", cioè ogni regola era costituita da una lista di condizioni che controllano se la regola è applicabile, e da una lista di azioni da compiere se la regola è applicabile.

Usando tali regole, nel corso delle sperimentazioni AM scoprì, ad esempio, i numeri primi. Ciò avvenne eseguendo il compito di trovare tutti i divisori di singoli numeri e classificare i risultati in insiemi.

Il programma trovò ben presto, con un'esplorazione empirica, che nessun numero aveva zero divisori e che soltanto uno ne aveva uno solo, ma trovò anche esempi di numeri con solo due divisori (cioè 1 e il numero stesso). In tal modo aggiunse alla sua raccolta di concetti quello di numero avente solo due divisori, che è il concetto di numero primo.

Incidentalmente, è molto interessante il fatto che AM abbia generato un altro concetto di teoria dei numeri, che era completamente nuovo per Lenat (e per me), ma che era stato studiato nei primi anni del secolo dal giovane genio indiano Ramanujan. È il concetto dei numeri "massimamente composti", che sono una specie di opposto del numero primo. Sono infatti quei numeri che hanno un numero di divisori maggiore dei loro inferiori. Così i primi numeri massimamente composti sono 1, perché ha solo 1 divisore, 2 perché ne ha 2, 4 perché è il primo che ne ha 3, 6 perché è il primo che ne ha 4, e così via.

AM non scoprì soltanto i numeri naturali e quelli primi (in più di un modo), ma elaborò anche il teorema di fattorizzazione unica, che dice che ogni numero naturale è scomponibile in un prodotto di numeri primi in un modo solo; inoltre scoprì l'enunciato di Goldbach (finora non dimostrato) che ogni numero pari è la somma di due numeri primi.

A simili notevoli successi si affiancano fallimenti altrettanto interessanti. Ho già accennato al fatto che il programma non seppe scoprire le frazioni. Non fece neppure grandi progressi nella teoria degli insiemi, anche se scoprì le relazioni più semplici come quelle di De Morgan. Seguì spesso percorsi inutili, occupandosi per esempio del concetto di numeri che possono essere rappresentati come la somma di due primi in un unico modo soltanto. Nella sua tesi, Lenat stimava che AM aveva scoperto circa 25 concetti buoni ("vincenti"), 100 accettabili e 60 "perdenti". Altrove [9], ho discusso possibili spiegazioni di questo comportamento; un fattore importante è, credo, l'esclusione dal suo modello di tentativi di risolvere particolari problemi matematici imposti dall'esterno. Simili tentativi sono stati una delle fonti maggiori di scoperta di nuovi concetti nello sviluppo storico della matematica, e, per esempio, un tentativo andato a vuoto di risolvere certe equazioni numeriche lineari avrebbe sicuramente condotto alla scoperta delle frazioni.

6. Modellistica cognitiva e linguaggi di programmazione

Ho fatto tre esempi di programmi relativamente recenti (per giocare a backgammon, per comprendere il linguaggio naturale, per fare scoperte matematiche) per mostrare non infondata la mia affermazione che programmi al livello superiore dello spettro che ho precedentemente discusso mostrano processi che non sono "meccanici" nel senso deteriore del termine, ma imitano quelle caratteristiche di creatività, intuizione, intelligenza, ecc. che appartengono alla psiche umana. Avrei potuto trovare esempi altrettanto convincenti in altri settori, ad es. diagnostica medica [10], percezione visiva [11], prova di proprietà dei programmi [12], test di quoziente di intelligenza [13], apprendimento [14], etc.; ma la domanda da fare è se da tutto questo lavoro e sperimentazione sono emersi dei principi generali.

Prima di rispondere a questa domanda dobbiamo delimitare il campo di ciò di cui parliamo. I processi mentali che sono stati modellati in questi programmi sono, parlando in generale, quelli "consci" o, secondo Freud, "preconsci", cioè processi piuttosto facilmente accessibili all'introspezione. Ma, come ha dimostrato il lavoro di Freud e dei suoi successori, una grande quantità e forse alcune delle parti più importanti delle nostre attività mentali operano a livello "inconscio", dove sono implicati i sentimenti, le emozioni, gli affetti. Personalmente non farei una netta distinzione tra processi "inconsci" e "preconsci", e perciò non escluderei la possibilità di trovare un modello computazionale, ad esempio, delle emozioni, per quanto strano ciò possa apparire a qualcuno. Vorrei soffermarmi più avanti su questo o meglio sulle basi di una tale possibilità. Per ora mi propongo di rispondere alla questione che ho sollevato, circa l'emergenza di principi generali, soltanto per i programmi di tipo più "conscio".

La mia risposta alla domanda sarebbe "si", sebbene i principi interessati siano di tipo differente da quelli delle scienze fisiche. Essi non sono quantitativi, bensì rappresentativi. Un esempio tipico di principio del primo tipo è quello "del minimo tempo", che stabilisce che in ogni sistema ottico il percorso seguito da un raggio di luce tra due punti qualsiasi è tale che il tempo impiegato risulta minore di quello che si avrebbe seguendo qualsiasi altro percorso fra gli stessi estremi. Con principi di questo genere non c'è da sorprendersi che gli strumenti principali della teoria e delle applicazioni di queste scienze siano branche tradizionali della matematica come l'algebra e l'analisi. Nelle scienze cognitive, invece, più che di misure numeriche, ci si occupa della rappresentazione di processi complessi, e perciò non ci si deve meravigliare che i mezzi principali di indagine siano i linguaggi, e in particolare quelli dei programmi degli elaboratori. Ci si deve quindi aspettare che i principi generali ai quali ho fatto riferimento possano essere scoperti nei tratti rappresentativi comuni dei linguaggi di programmazione che sono stati e sono tuttora sviluppati per l'intelligenza artificiale. Vorrei parlarne nell'ordine approssimativo di sviluppo cronologico sotto le voci seguenti: strutturazione dei simboli, ricerca, obiettivo, pattern-matching e cornice.

Strutturazione dei simboli

Poiché ci si occupa di conoscenze di ogni genere piuttosto che di operazioni di aritmetica, occorre un modo facile e flessibile di scrivere ogni tipo di struttura simboli-

ca richiesta. Questa necessità fu risolta dai linguaggi orientati alle liste sviluppati negli anni '50 e '60 come IPL.V a Pittsburg, LISP a Cambridge, Boston e POP a Edinburgh. Questi linguaggi usavano elenchi "annidati" di simboli, in cui le operazioni fondamentali erano la formazione di elenchi e la selezione di simboli dagli elenchi.

Ricerca (Search)

Come si è visto dagli esempi che ho dato, ci può essere maggiore o minore ricerca nel procedimento. È necessario che la rappresentazione includa sufficienti possibilità a questo riguardo.

Obiettivo

La maggior parte dei processi mentali, perlomeno del tipo "conscio", ma probabilmente anche quelli del tipo "inconscio", sono motivati da propositi; quindi devono essere rappresentati gli obiettivi che dirigono i processi, compresa la generazione di scopi e sottoscopi come parti del processo.

Pattern-matching

I primi programmi di intelligenza artificiale erano scritti in linguaggi che, essendo orientati verso un processo simbolico piuttosto che matematico, erano molto differenti dai linguaggi "ordinari" destinati al calcolo aritmetico, ma tuttavia gli assomigliavano, essendo "deterministici" nel senso che chi scriveva il programma doveva in effetti specificare l'ordine in cui le istruzioni o subroutines dovevano essere eseguite dall'interprete o dal compilatore. Così, ad esempio, se veniva chiamata una subroutine, essa poteva essere soltanto una di quelle nominate nel programma. Per ragioni di flessibilità e generalità, è richiesta la capacità di scegliere una subroutine adatta alla specifica necessità. A questo scopo, delle variabili vengono usate nel modo illustrato dal seguente esempio. Supponiamo che in un punto di un programma scritto per comprendere storie di bambini, il programma deve scoprire se Marco va a scuola. Tra i moltissimi dati immagazzinati nella base di dati, potrebbe avere i due seguenti (scritti in LISP):

.....

(IMPLIES (BOY X) (GOSCHOOL X))

.....

(BOY MARCO)

Ora il programma dovrebbe prima interrogare la base di dati su un argomento (GOSCHOOL MARCO). Se non lo trova, dovrebbe cercare se l'ultima sottolista in una lista avente IMPLIES come primo elemento, inizia con GOSCHOOL, cercando di combinare tale sottolista con l'obiettivo (GOSCHOOL MARCO). In questo caso, il programma riesce a identificare la variabile X con MARCO, generando così come sottobiettivo la ricerca nella base di dati di una voce (BOY MARCO), ottenendo così l'obiettivo originario di riconoscere che Marco andò a scuola.

Capacità di orientamento all'obiettivo, di ricerca e di pattern-matching furono per la prima volta incorporati nel linguaggio MICROPLANNER [15], a Cambridge, (Boston) e, successivamente nel linguaggio tipo PROLOG, a Marsiglia [16].

Cornice

La nozione di "cornice" fu esplicitamente introdotta da Marvin Minsky [17], anche se qui io userò il termine in un senso un po' più ampio dell'originale. Essa era un paradigma per ridurre a modello alcuni degli aspetti più "globali" di conoscenza che sembra siano responsabili della apparente velocità e immediatezza del nostro pensiero ordinario. Secondo Minsky, quest'ultimo, ben lontano dall'essere un qualche tipo di ricerca entro uno spazio di possibilità poco strutturato, procede in modo molto più diretto, facendo uso di "pezzi" di conoscenza di grosse dimensioni, pronti, ben organizzati, che trattano le situazioni sulla base delle aspettative. Prendete il suo esempio di come si vede una stanza. Immediatamente prima di entrare in una stanza ne sapete solitamente abbastanza per aspettarvi una stanza piuttosto che, diciamo, un paesaggio. Lo potete indovinare dal tipo della porta. Il riconoscimento della porta fa scattare una "cornice" di stanza. Se non è una certa stanza particolare che ci si aspetta, la "cornice" potrebbe contenere una rappresentazione della struttura schizzata in fig. 1.

Se è una stanza particolare che ci si aspetta di vedere, la "cornice" potrebbe contenere altri elementi, ad esempio un pianoforte nell'angolo in fondo a destra.

Per elaborare i dati visuali di ingresso, il sistema cognitivo ha un punto di partenza nel fatto che può adattare immediatamente le rappresentazioni (generate al proprio interno) di linea, vertice, area, etc. alle parti appropriate della cornice. Per di più, la cornice stessa può dirigere l'elaborazione, selezionando sia che cosa trattare sia quanto. Per esempio, avendo stabilito che c'è un lato diritto corrispondente ad A-B nella struttura, non occorre guardare B-C. Minsky riassume il suo concetto di "cornice" nel mòdo seguente:

"Quando si incontra una situazione nuova (o si ha un cambiamento sostanziale nella visione della situazione in atto), si sceglie dalla memoria una struttura sostanziale. Questa è una "cornice" registrata, che deve essere adattata alla realtà cambiando i dettagli secondo necessità".

Le strutture dei dati che sono state impiegate per realizzare "cornici", possono prendere forme differenti ed avere nomi diversi nella letteratura; io qui mi limiterò a menzionare un esempio dell'applicazione di questo potente principio. Molti, se non la maggior parte, dei programmi di linguaggio naturale del gruppo di Schank ai quali ho fatto riferimento prima, usano i cosiddetti "scripts" che descrivono in dettaglio gli eventi che danno luogo ad una situazione standard. Così uno script di ristorante è costituito da tutti gli eventi che normalmente si verificano nel cenare al ristorante, insieme alle opzioni su che cosa fare se qualcosa va storto, e ad elementi che dipendono dal tipo di ristorante in cui vi trovate.

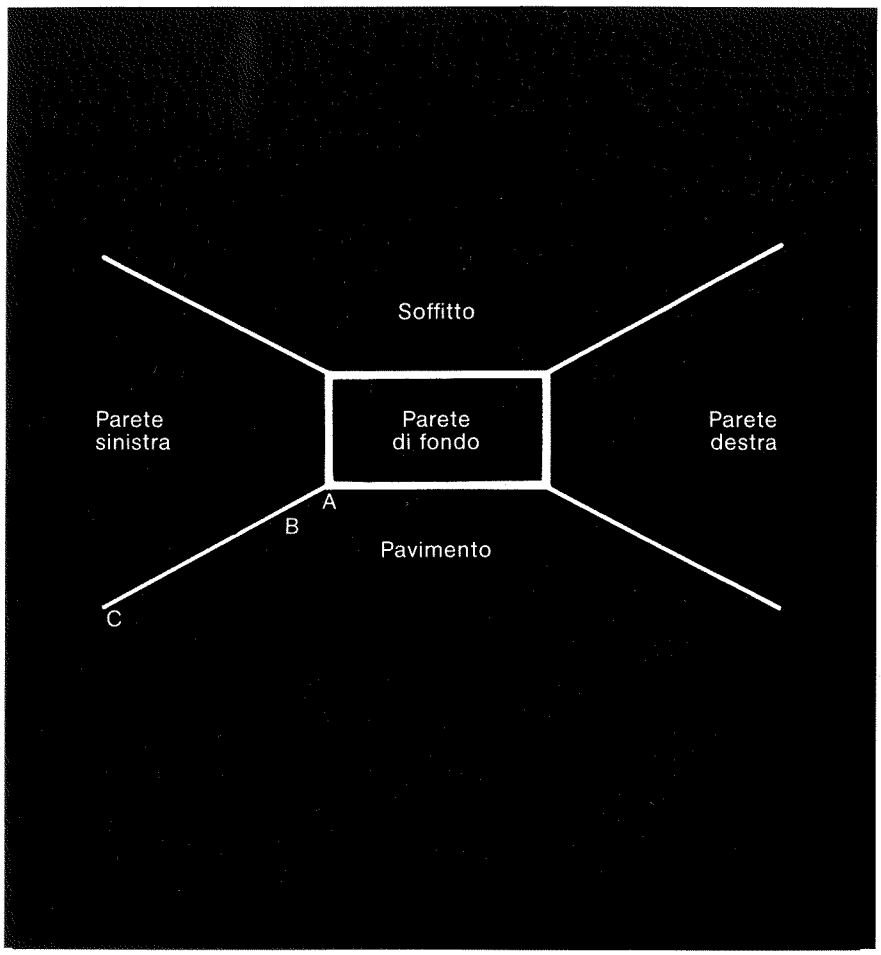


Fig. 1 - Il concetto di "cornice„ relativo ad una stanza.

7. Speculazioni su un nuovo paradigma per la psicologia

Tutto ciò che ho detto sin qui e gli esempi che ho portato, mi sembra che implichino la convinzione che i programmi degli elaboratori sono in grado di fare modelli di processi mentali complessi e di suggerire principi generali che governano i processi mentali, come nessun altro approccio conosciuto in psicologia è in grado di fare. Vorrei ora sviluppare alcune considerazioni su una teoria unificata dei processi mentali, del tipo sia "conscio" che "inconscio", che è suggerita non soltanto dalla nostra esperienza nel fare modelli computazionali, ma anche dai risultati sperimentali della neurofisiologia degli ultimi decenni e da un recente contributo alla filosofia del linguaggio.

Anzitutto, vorrei riprendere un punto cui ho accennato in precedenza, quando ho fatto rilevare che i programmi cognitivi che ho citato, quali quelli di Berliner o Schank, hanno a che fare esclusivamente con processi mentali "consci" o "preconsci", e non con processi di tipo "inconscio". È possibile che abbia confuso la natura dei compiti da svolgere con i processi per effettuarli. Giocare a scacchi e fare riassunti dei giornali sono certamente compiti "consci", ma i processi coinvolti nel farli non necessariamente sono del tipo che riconosciamo come "consci" o "preconsci". Per esempio, le rappresentazioni del significato delle parole nei programmi con "dipendenza concettuale" di Schank sono tutte espresse in termini di un ristretto numero di cosiddette "primitive semantiche". Come esempio, si può citare la primitiva PTRANS (questo nome è probabilmente l'acronimo di "Transfer of physical matter"), che costituisce il nucleo della struttura dei dati per parole come "mangiare", il cui significato implica il movimento di materia da una parte ad un altra, in questo caso di cibo dall'esterno alla bocca. In sostanza, la teoria della comprensione del linguaggio naturale implicita nel modello di Schank sostiene che nei nostri processi mentali esiste attualmente la rappresentazione di, diciamo, circa dieci di tali primitive, che vengono usate quando abbiamo a che fare con parole.

Ma certamente noi non siamo consci di un tale uso, ed io credo che sarebbe forzare eccessivamente il significato del termine se diciamo che esse sono "preconscie".

La mia opinione è che in un tale modello computazionale sono inclusi processi apparentemente "inconsci".

Un altro punto che vorrei considerare riguarda un aspetto della natura del linguaggio. La scienza degli elaboratori ci ha famigliarizzati con linguaggi aventi lo scopo non di comunicare, ma di rappresentare e di elaborare, per esempio linguaggi assemblativi e linguaggi di macchina. Recentemente, una stimolante analisi filosofica di Aaron Sloman [18] suggerisce che questa

è in effetti la principale funzione del linguaggio, la funzione di comunicazione essendo solo un "prodotto derivato". Vale a dire che gli animali ed i nostri lontani antenati umani avevano già operanti dentro di sè linguaggi di rappresentazione ed elaborazione, i quali, nel corso dell'evoluzione, vennero, per così dire, portati all'aperto attraverso l'invenzione di segni esterni, sotto forma di suoni e di immagini. Si potrebbe obbiettare che questo è forzare il significato del termine "linguaggio", poiché, dopo tutto, ciò che presumibilmente si svolge nel nostro corpo sono interazioni neuroniche di vario tipo o cose del genere: dove sono le parole, le frasi, ecc.? Questo sarebbe, però, voler adottare un irrealistico punto di vista sul linguaggio. Consideriamo che cosa accade quando io vi parlo. Un processo al mio interno dà luogo alla emissione di perturbazioni dell'aria, che viaggiano sino alle vostre orecchie e danno luogo ad un processo dentro di voi. Come dobbiamo considerare il fenomeno fisico di queste perturbazioni dell'aria che viaggiano tra me e voi? Esso *è* un campione di linguaggio, e non solo perché voi lo avete interpretato come tale. Anche se nessuno lo avesse udito, o lo udrà mai, esso sarebbe pur sempre linguaggio; infatti qualcuno potrebbe avere avuto un dispositivo rivelatore-interprete che lo avrebbe decodificato. (Allo stesso modo, i segni stampati su un libro sono linguaggio anche se nessuno lo leggerà mai). La perturbazione dell'aria trasportava una rappresentazione che avrebbe innescato altri processi nel corpo dell'ascoltatore. Se dunque un fenomeno meccanico così banale come le vibrazioni dell'aria può trasportare il linguaggio, perché non dovrebbero gli eventi neuronici fare lo stesso negli organismi biologici?

Ci sono ormai dei risultati ottenuti da suggestivi esperimenti in neurofisiologia negli ultimi anni, che vanno in questa direzione. Per esempio, Barlow [19] riferisce che lo studio per registrare la risposta di cellule con un singolo nervo nel sistema visivo di animali come pesci e scimmie, ha mostrato che l'occhio usa un alfabeto di solo una o due dozzine di simboli. Questi simboli sono generati da fattori "grilletto", quali la comparsa nel campo visivo di una piccola macchia nera, o il movimento verso l'alto di un oggetto in una particolare regione. È quindi evidente che gli eventi in queste celle nervose agiscono essenzialmente come le perturbazioni dell'aria nel parlato umano: essi portano rappresentazioni ed eccitano eventi uguali o similari negli organismi.

Per tornare alla psicologia umana, anche se si accetta l'esistenza di un linguaggio interno di elaborazione che si sostanzia in eventi del sistema nervoso, non sarebbe più sicuro e più in accordo con quanto conosciamo tracciare una distinzione netta tra quei linguaggi interni e ciò che, per il momento, vorrei indicare come i linguaggi esterni che noi parliamo, quale l'inglese o l'italiano? Io credo di no, sia per ragioni di evidenza sia per il potenziale di chiarimenti implicito in una visione unitaria del fenomeno.

Mi riferisco alla suggestiva evidenza derivante dall'apprendimento umano, specialmente quando viene coinvolto l'uso di simbolismi piuttosto formali. Prendete come un esempio della stretta associazione e continuità tra linguaggio interno ed esterno i cambiamenti che avvengono nel nostro apprendimento dell'aritmetica elementare durante l'infanzia. Quando ci insegnano per la prima volta la somma, noi la eseguiamo in termini di istruzioni del maestro, usando consciamente parole e frasi del linguaggio naturale come "riporto" e "due volte tre fa sei"; ma quando siamo diventati abili ed eseguiamo le somme nel modo usuale, automatico, "senza pensare", non rimane praticamente nulla nella nostra coscienza di quelle istruzioni ed espressioni. Ma nessuno può dubitare, sebbene non ne siamo consci, che sia in atto una elaborazione simbolica: l'attività è passata sostanzialmente dal linguaggio esterno a quello interno. Si noti quanto facile e "naturale" sia la transizione tra i due linguaggi: se, allorché siamo diventati abili, ci troviamo di fronte ad una somma particolarmente lunga o difficile, noi facciamo talvolta la transizione inversa, ritornando alla parte esterna del linguaggio, dicendo magari a noi stessi: "Devo ricordare di riportare 2 quando ho finito di sommare questa colonna".

L'ipotesi che pertanto io vorrei fare è la seguente:
"Tutti i processi mentali in qualsiasi sistema biologico sono pure trasformazioni che hanno luogo all'interno del linguaggio proprio del sistema, e questo linguaggio può essere attivato ad ogni livello".

La mia opinione è che questo punto di vista è potenzialmente capace di fare luce su un ampio numero di temi in psicologia (così come in linguistica). Nella parte finale di questo articolo passerò brevemente in rassegna alcuni di questi temi, ma prima vorrei chiarire che cosa intendo per "livello" nella mia definizione, facendo una analogia con l'elaboratore.

L'idea di livello è ben nota e chiara in un sistema di elaborazione. Un programma scritto nel linguaggio sorgente al più alto livello, per essere eseguito deve essere prima convertito nel linguaggio assemblativo della macchina ed il risultato convertito successivamente in codice di macchina. Uno schema gerarchico stratificato di questo tipo è il modo normale con cui i programmi vengono eseguiti. Tuttavia sono possibili ampie variazioni; per esempio, alcuni sistemi offrono la possibilità (generalmente per ragioni di efficienza) di interpolare pezzi di programma in codice macchina nel programma scritto in linguaggio al più alto livello. Quando si fanno cose di questo tipo, o anche semplicemente si medita la possibilità di farle, allora la nozione di strati di linguaggi diversi nella macchina comincia a perdere senso, e può diventare più utile pensare al programma come scritto in un unico linguaggio (di cui il sorgente, l'assemblativo e quello in codice-

macchina, così come le trasformazioni tra essi, sono i costituenti). Questo può essere l'analogo del linguaggio unitario degli organismi biologici che avevo postulato. E quando parlavo di "livelli" ai quali esso può essere attivato, mi riferivo puramente a delle posizioni in questo complesso sistema linguistico.

8. Alcune implicazioni dell'approccio unitario

Ho discusso questo argomento altrove [20], ma vorrei richiamare qui alcune delle conclusioni sotto differenti titoli.

Interazione di modi cognitivi

Il fatto che, per esempio, noi possiamo immaginare degli oggetti senza vederli ma soltanto sentendo il loro nome, richiede una spiegazione. Questo fatto risulta spiegabile nell'ipotesi che il funzionamento delle nostre differenti modalità cognitive e sensorie abbia un linguaggio comune, per cui le stesse rappresentazioni possono essere raggiunte attraverso differenti percorsi di elaborazione linguistica. Questo potrebbe essere un esempio di attivazione del linguaggio unitario a differenti "livelli", di cui ho in precedenza parlato.

Attività artistiche

La poesia è particolarmente interessante da considerare in quanto, mentre usa esclusivamente linguaggio naturale, sembra – proprio perché si tratta di poesia e non di prosa – che cerchi di evocare "livelli più profondi" della parte interna del linguaggio. La poesia attiva risonanze che il normale uso del linguaggio non sembra in grado di suscitare. È noto che non di rado la poesia più efficace presenta un piccolo significato discernibile in termini di discorso ordinario, e tuttavia essa ha chiaramente un grande significato in termini di processo mentale. Sembra che la poesia cerchi di ricreare, di rispecchiare, di rappresentare in linguaggio naturale, delle parti estremamente significative del nostro linguaggio interno.

Anche la musica non utilizza il linguaggio discorsivo e pertanto, non dovendo agire attraverso tale canale, appare adatta a suscitare una più diretta, e spesso potente, evocazione del nostro linguaggio interno.

Per quanto concerne le arti figurative, secondo questa ipotesi, probabilmente i nostri antichi antenati trasferirono sulle pareti delle loro caverne parti selezionate della loro interna rappresentazione del mondo; e forse altrettanto hanno fatto da allora tutti gli artisti. È significativo che nell'importante lavoro di David Marr sui modelli computazionali della percezione visiva di oggetti [11], uno degli stadi di elaborazione previsti è quello che egli chiama "stick figures", caratteristica di uno degli stili di Picasso.

Emozioni

I miei commenti su questo aspetto sono ancora più sommari e di tentativo che per gli aspetti precedenti. È mia impressione che non si siano fatti molti progressi in questa area della psicologia. Se ci rifacciamo alla teoria delle emozioni di James Lange ed al lavoro di Walter Cannon e della sua scuola, si può dire che parecchio è stato scoperto sulle connessioni tra fenomeni emotivi e specifiche funzioni corporee, come quelle del sistema endocrino. D'altra parte c'è stata anche, se pur in grado minore, una certa attenzione agli aspetti cognitivi. Circa questi ultimi, vorrei citare un breve estratto da uno scritto particolarmente interessante di Elizabeth Duffy [21]:

> "Io credo che si possa essere d'accordo sul fatto che un individuo non sente "emozione" se non in situazioni che hanno significato per lui. Egli è "abbattuto" o "irato"... quando è bloccato nel raggiungimento di un importante obiettivo. Egli è invece "lieto"... allorché viene facilitato nel raggiungere tale obiettivo. Il livello o l'intensità dell'emozione è grossolanamente proporzionale al grado di importanza del particolare obiettivo..."

Poiché abbiamo visto che i metodi dell'intelligenza artificiale bene si adattano alla rappresentazione di obiettivi, aspettative e interpretazioni, non dovrebbe essere troppo difficile determinare efficaci modelli delle emozioni.

Psicologia dell'incoscio

Il classico lavoro pionieristico di Freud sui processi inconsci che intervengono nei sogni, negli scherzi e nelle parafrasi [22, 23, 24], sembra ad un profano, come lo scrivente, aver avuto assai minor seguito scientifico di quanto ci si sarebbe potuto aspettare. È possibile che ciò sia dovuto non solo al molto maggiore interesse suscitato nelle applicazioni mediche che in quelle scientifiche dalle idee e dalle tecniche di Freud, ma anche al fatto che ai suoi tempi, e per parecchio tempo dopo, non sono esistiti veicoli adeguati per la formulazione di teorie che fossero sia precise sia adatte a verifiche. Ho già fatto rilevare come le formulazioni della matematica algebrica e analitica, al pari di quelle della fisica, siano del tutto inadatte ai processi mentali. Se ora consideriamo lo stretto determinismo della visione freudiana degli eventi mentali, la sua grande scoperta della tecnica di associazione libera per ottenere l'accesso all'inconscio, l'affascinante intreccio di sintassi e semantica delle spiegazioni, i tipi di trasformazione, da lui isolati e studiati, nelle transazioni dell'inconscio, quali trasposizione, condensazione e inversione; allora ci sembra di poter dire che un linguaggio unitario di elaborazione dei processi mentali possa essere il mezzo giusto per una ulteriore appropriata formulazione della teoria e per il suo sviluppo. Mi limiterò qui a dare un'idea di alcune possibilità che si potrebbero aprire.

I fenomeni di base dell'ipnotismo potrebbero essere interpretati come il trasferimento di comandi dalla parte esterna a quella interna del linguaggio.

Le trasformazioni, cui ho accennato, che hanno luogo nei sogni e negli scherzi, potrebbero collocarsi, con una formulazione altrettanto

rigorosa, nell'ambito di trasformazioni informatiche già note, quali l'interpretazione, la compilazione ed altre ancora usate nella modellistica computazionale. Non dovrebbe allora essere troppo difficile scrivere, per esempio, un programma che, con un dato corpo di conoscenze, possa generare ottime facezie!

La tecnica della libera associazione potrebbe essere usata per illuminare i costituenti inconsci dei fenomeni mentali più normali, oltre che di quelli patologici che interessano gli psicoanalisti. Lo stesso processo di associazione libera potrebbe essere suscettibile di modellazione computazionale e pertanto di studio sperimentale. Incidentalmente, una delle ironie dialettiche della teoria di Freud è che l'importanza della libera associazione risiede nel fatto che essa *non* è libera; essendo essa strettamente determinata dalla psiche, proprio per questo è capace di fornire indizi sulla struttura di quest'ultima.

Infine, sebbene abbia evitato di discutere di psicopatologia in senso stretto, vorrei attirare l'attenzione sul fatto, rilevato sia da Freud che dal poeta e pensatore francese Paul Valéry, che gli effetti di alcune delle nostre più traumatiche esperienze psichiche possono essere paragonati a quelli provocati da un corto-circuito in un sistema elettrico complesso, quale l'appiccare fuoco ad una casa. Ai loro tempi, essi dovevano ricorrere ad analogie fisiche, ma in futuro, forse, si scoprirà che molte condizioni psicopatologiche trovano la loro origine in relativamente piccoli malfunzionamenti dei nostri linguaggi interni di elaborazione.

9. Conclusioni

Ho già esposto altrove [20] le possibili interessanti implicazioni della tesi unitaria sulla struttura e la semantica dei linguaggi naturali. Vorrei quindi concludere qui questo articolo facendo rilevare che esso è costituito da due parti. La prima parte è sostanzialmente il resoconto fattuale di alcuni sviluppi della scienza degli elaboratori, e quindi va giudicata come tale. La seconda parte è costituita invece da pure speculazioni sugli sviluppi cui può portare la prima, e quindi deve essere giudicata in una luce differente.

10. Bibliografia

[1] FREUD S. (1962) "The question of lay analysis" in *Two Short Accounts of Psychoanalysis* (Harmondsworth: Penguin Books) p. 103.

[2] BERLINER H. (1980) "Computer backgammon", *Scientific American*, vol. 242, No. 6, pp. 54-62.

[3] BERLINER H. (1979) "On the construction of evaluation functions for large domains" in *Proceedings of the Sixth International Joint Conference on Artificial Intelligence*, pp. 53-55.

[4] SAMUEL A.L. (1959) "Some studies in machine learning using the game of checkers", *IBM Journal of Research and Development*, vol. 3, No. 2, pp. 210-229.

[5] SCHANK R.C. (1979) "Natural language, philosophy, and artificial intelligence", in Ringle M.D. (ed.) *Philosophical Perspectives in Artificial Intelligence* (Brighton: Harvester Press), pp. 196-224.

[6] SCHANK R.C. (1975) *Conceptual Information Processing* (Amsterdam: North-Holland).

[7] LENAT D.B. (1976) *AM: An Artificial Intelligence Approach to Discovery in Mathematics as Heuristic Search* (Ph.D Thesis, University of Stanford).

[8] LENAT D.B. (1977) "The ubiquity of discovery", *Artificial Intelligence*, vol. 9, No. 3, pp. 257-288.

[9] MELTZER B. (1977) Review of Lenat's Ph.D. thesis *AISB Quarterly* issue 27, pp. 20-23.

[10] SHORTLIFFE H.S. (1976) *Computer-based Medical Consultations: MYCIN* (New York: Elsevier).

[11] MARR D. (1979) "Visual information processing", in *Proceedings of the Sixth International Joint Conference on Artificial Intelligence*, pp. 1108-1126.

[12] BOYER R.S. and MOORE J. (1979) *A Computational Logic* (New York: Academic Press).

[13] EVANS T. (1964) "A program for the solution of geometric-analogy intelligence questions", in Minsky M. (ed.) 1968, *Semantic Information Processing* (Cambridge, Boston: MIT Press), pp. 271-353.

[14] WINSTON P.H. (1970) *Learning Structural Description from Examples*, Technical Report AI TR-231 (Cambridge, Boston: Artificial Intelligence Laboratory, MIT).

[15] SUSSMAN G., WINOGRAD T. and CHARNIAK E. (1970) *Micro-planner Reference Manual* (Cambridge, Boston: Artificial Intelligence Laboratory, MIY).

[16] ROUSSEL P. (1975) *PROLOG: Manuel de Reference et d'Utilisation* (Luminy: Groupe d'Intelligence Artificielle, Université d'Aix-Marseille).

[17] MINSKY M. (1975) "A framework for representing knowledge", in Winston P.H. (ed.) *The Psychology of Computer Vision* (New York: McGraw-Hill), pp. 211-277.

[18] SLOMAN A. (1979) "The primacy of non-communicative language", in *Proceedings of the 5th Informatics ASLIB Conference*, Oxford, 26-28 March 1979.

[19] BARLOW H.B. (1977) "The languages of the brain", in Duncan R. and Weston-Smith M. (eds.) *The Encyclopaedia of Ignorance, Vol. 2*, (Oxford: Pergamon Press), pp. 259-272.

[20] MELTZER B. (1979) "Some speculations on language", in *Proceedings of the Sixth International Joint Conference on Artificial Intelligence*, pp. 1127-1130.

[21] DUFFY E. (1941) "An explanation of 'emotional' phenomena without the use of the concept 'emotion'", *Journal of General Psychology*, vol. 25, pp. 283-293.

[22] FREUD S. (1978) *The Interpretation of Dreams* (Harmondsworth: Penguin Books).

[23] FREUD S. (1978) *Jokes and Their Relation to the Unconscious* (Harmondsworth: Penguin Books).

[24] FREUD S. (1978) *The Psychopathology of Everyday Life* (Harmondsworth: Penguin Books).

Ricerca documentaria e automazione

RANIERI TESI
Honeywell Information Systems Italia
Servizio Ricerche e Sviluppo Marketing
Milano

1. Introduzione - Il problema dell'information retrieval

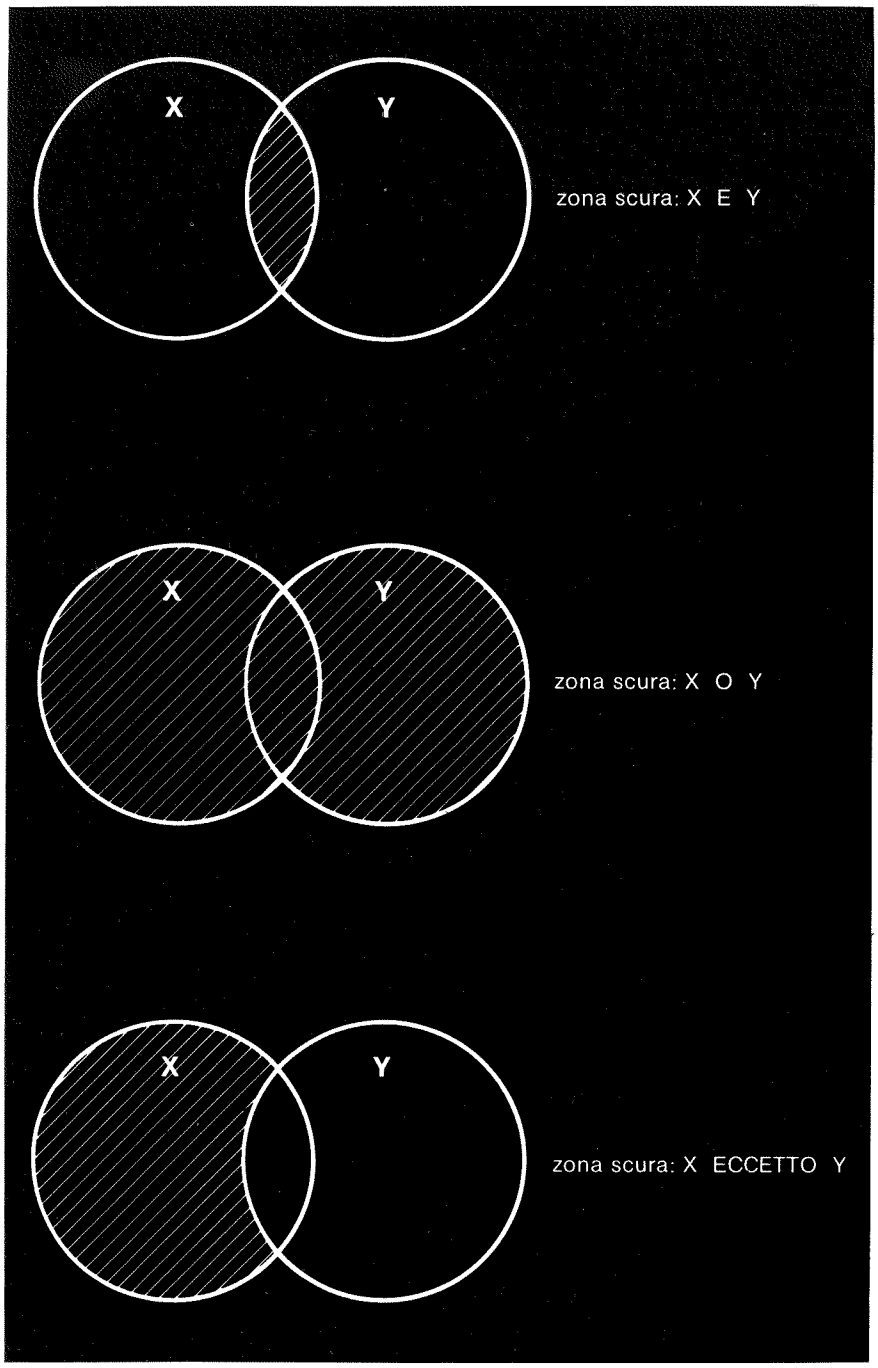
Pur senza la pretesa di affrontare un problema di classificazione generale dell'informazione, si possono tuttavia distinguere due grandi categorie di "messaggi", che potremmo definire rispettivamente "messaggi diretti" e "messaggi indiretti". I messaggi diretti sono caratterizzati dal fatto di essere recepiti e utilizzati dai destinatari nel momento stesso della loro diffusione. I messaggi indiretti debbono essere resi disponibili all'utente in modo tale che quest'ultimo possa prenderne visione quando lo ritiene più opportuno. La radio, la televisione, l'altoparlante di una stazione ferroviaria offrono altrettanti esempi di informazione diretta, ossia di un'informazione che raggiunge il suo scopo principale e più immediato non appena raggiunto il destinatario. Libri, riviste, testi di legge e, in generale, ogni documentazione scritta rappresentano al contrario casi di informazione indiretta, cioè di un'informazione che, per sua natura, necessita di essere resa accessibile all'utente nel corso di un arco di tempo più o meno lungo. I problemi tecnico-organizzativi posti dalla circolazione dell'informazione si diversificano radicalmente a seconda che si tratti di informazione diretta o indiretta. Nel primo caso sono prioritari i problemi di diffusione, nel secondo quelli di archiviazione. Questo articolo è dedicato in particolare alla ricerca documentaria, che costituisce la più attuale proposta di soluzione del problema dell'accesso agli archivi.

L'informazione indiretta, dovendo risultare accessibile per periodi più o meno lunghi, dev'essere "archiviata", ossia "concentrata" in luoghi presso i quali l'utente possa rivolgersi per ottenere l'informazione desiderata. Chiameremo tali luoghi "archivi". Esempi di archivi sono le biblioteche, gli archivi di stato, le raccolte di documenti aziendali ecc. L'archivio rappresenta lo strumento naturale per risolvere il problema dell'ubicazione dell'informazione indiretta, ma ne pone uno nuovo: quello della ricerca dell'informazione al suo interno. Nelle organizzazioni "tradizionali", tale ricerca avviene utilizzando uno o più indici. L'indice, per quanto possa essere definito con cura, trova tuttavia un limite alla propria efficacia nell'ipotesi sulla quale si basa: la conoscenza da parte dell'utilizzatore di precisi attributi dei documenti che cerca. Per poter usare un indice per argomento, poniamo, l'utilizzatore deve non solo conoscere l'oggetto di ciò che cerca, ma deve anche essere in grado di qualificarlo con gli stessi termini impiegati nell'indice. Per poter usare un indice per autore, ad esempio, l'utente deve conoscere il nome esatto di quest'ultimo. Come si vede con chiarezza da questi esempi, la ricerca diviene problematica ogni volta che l'utente non sia in grado di "descrivere" ciò che intende trovare con termini eguali a quelli usati nella formulazione delle voci d'indice. I limiti dei mezzi "tradizionali" di ricerca aumentano, naturalmente, in rapporto ai volumi e alle dimensioni degli archivi, fino a renderli, in taluni casi, inaccessibili se non da parte di specialisti. Le tecniche di ricerca documentaria (information retrieval), basate sugli elaboratori elettronici, si pongono come obiettivo la rimozione o almeno l'allargamento di tali limiti. Il sistema "ideale" di information retrieval si può infatti definire come un sistema capace di individuare un qualsiasi insieme di documenti "qualunque" sia la "descrizione" fornita in input dall'utente, purché questa, beninteso, sia sufficiente in sè a qualificare l'oggetto della ricerca. È chiaro che nessun sistema "reale" di information retrieval realizza in pieno l'obiettivo "ideale", tuttavia è indubbio che anche in questo settore applicativo siano stati compiuti negli ultimi anni notevoli progressi.

2. Funzionalità fondamentali di un sistema di ricerca documentaria

I sistemi di ricerca documentaria, indipendentemente dalle modalità tecniche della loro realizzazio-

Fig. 1 - Operatori logici booleani.



ne, si basano su due funzionalità fondamentali: la *ricerca primaria*, o ricerca tramite flusso inverso, e la *ricerca secondaria*, o ricerca su testo.

La ricerca primaria. Questo tipo di ricerca consiste praticamente nella consultazione di un indice "onnicomprensivo" chiamato *flusso inverso*, e può essere realizzato attraverso le seguenti fasi:

1) Il sistema registra i documenti, interi o riassunti, su un opportuno supporto, ad esempio disco magnetico. L'insieme dei documenti registrati prende il nome di *flusso diretto*.
2) Il sistema "analizza" ogni documento e ne estrae, secondo criteri che vedremo nel seguito, un insieme di *parole-chiave* (o *descrittori*), ovvero un insieme di parole, locuzioni o codici, ciascuna delle quali rappresenta una caratterizzazione notevole del documento analizzato.
3) Il sistema compone, con le parole-chiave estratte dai documenti, una tabella, detta *flusso inverso*, che associa a ciascuna parola-chiave i documenti dai quali è stata estratta. Ad esempio, supponiamo che il flusso diretto si riferisca ad una biblioteca e sia composto dagli estratti di quattro volumi:

Vol. 1: Matematica elementare - Autore: Rossi
Vol. 2: Matematica e Fisica - Autore: Bianchi
Vol. 3: Matematica e Statistica - Autore: Verdi
Vol. 4: Storia della Matematica - Autore: Bianchi

Se le parole-chiave individuate sono:

Vol. 1: Matematica, Matematica elementare, Aritmetica, Rossi
Vol. 2: Matematica, Fisica, Bianchi

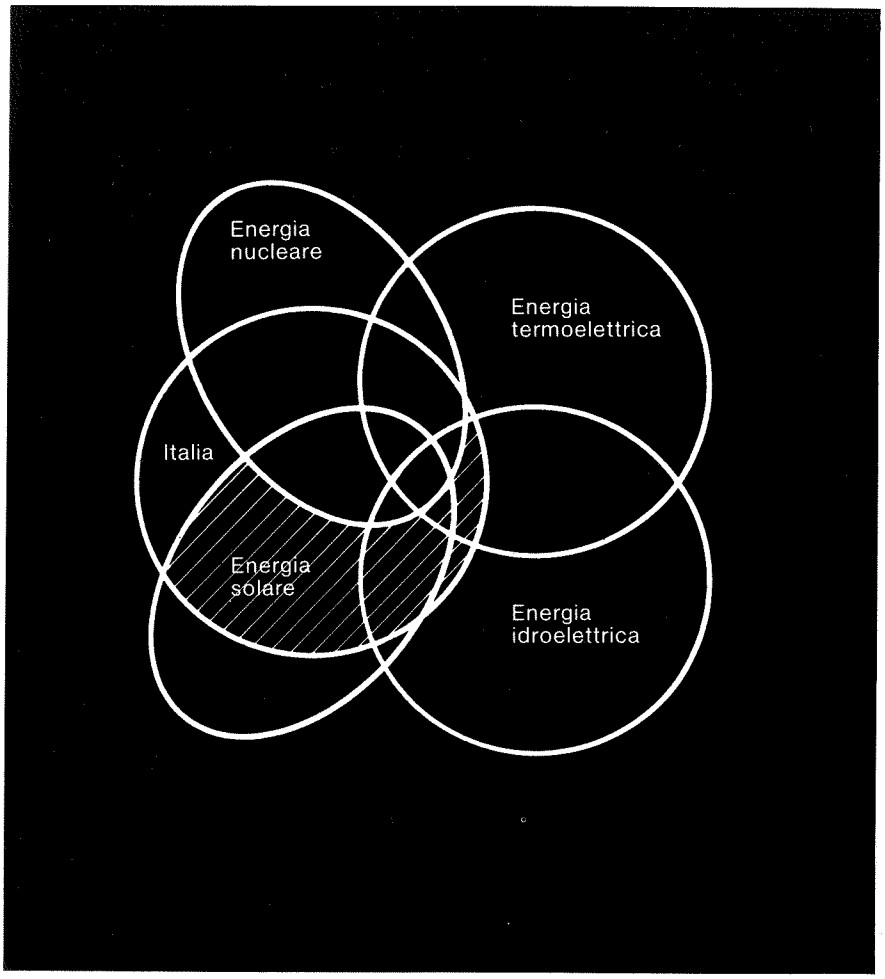


Fig. 2 - Esempio di applicazione degli operatori logici booleani. Ricerca di una bibliografia sulle disponibilità di energia non nucleare in Italia. La domanda si formula nel modo seguente: "Italia E (energia idroelettrica O energia termoelettrica O energia solare) ECCETTO energia nucleare".
La zona tratteggiata rappresenta il risultato della domanda.

Vol. 3: Matematica, Statistica, Verdi
Vol. 4: Matematica, Storia, Greci, Egiziani, Arabi, Bianchi

il corrispondente flusso inverso sarà:

Arabi: vol. 4
Aritmetica: vol. 1
Bianchi: vol. 2, vol. 4
Egiziani: vol. 4
Fisica: vol. 2
Greci: vol. 4
Matematica: vol. 1, vol. 2, vol. 3, vol. 4
Matematica elementare: vol. 1
Rossi: vol. 1
Statistica: vol. 3
Storia: vol. 4
Verdi: vol. 3

4) L'utente "interroga" il sistema fornendo in input le parole-chiave che "descrivono" l'oggetto della sua ricerca e il sistema "risponde" esplorando il flusso inverso e individuando in questo modo i documenti di provenienza delle parole-chiave. Continuando l'esempio sviluppato al precedente punto 3):

Domanda: Matematica
Risposta: vol. 1, vol. 2, vol. 3, vol. 4

Domanda: Rossi
Risposta: vol. 1

Domanda: Bianchi
Risposta: vol. 2, vol. 4

Domanda: Matematica elementare
Risposta: vol. 1

Per estendere le possibilità di interrogazione, si prevede inoltre l'impiego, nelle domande, di *operatori logici booleani* (v. figg. 1 e 2). Gli operatori logici booleani, variamente definiti a seconda delle concrete modalità di realizzazione del sistema, si possono comunque sempre ricondurre a tre fondamentali operazioni: E (intersezione), O (unione), NON (esclusione, ossia passaggio al complementare). Con riferimento all'esempio precedente, il loro funzionamento può essere così illustrato:

- Domanda: Matematica E Rossi
  Risposta: vol. 1

in questo caso l'utente ha chiesto i documenti in cui compaiono entrambe le parole-chiave Matematica e Rossi.

- Domanda: Fisica O Statistica
  Risposta: vol. 2, vol. 3

in questo caso l'utente ha chiesto i documenti in cui compaia almeno una delle parole-chiave Fisica e Statistica.

- Domanda: Matematica NON Fisica
  Risposta: vol. 1, vol. 3, vol. 4

con questa domanda l'utente ha chiesto i documenti in cui compaia la parola-chiave Matematica ma dai quali sia eslusa la parola-chiave Fisica.

La ricerca secondaria. Questa seconda funzionalità di ricerca permette di individuare i documenti che rispondono a determinate caratteristiche o condizioni espresse

dalle "domande" dell'utente.

Il sistema effettua la ricerca dei documenti desiderati analizzandone i testi, parola per parola o carattere per carattere, senza fare ricorso al flusso inverso. Le condizioni che si possono richiedere sono di vario tipo e ne diamo qui di seguito alcuni esempi significativi:

- presenza di una parola data in una data parte del documento; la "risposta" alla "domanda" che esprime questa condizione consiste nei documenti che contengono la parola data nella parte data (la parte può essere, ad sempio, il titolo o il riassunto); si noti che la parola ricercata può non essere una parola-chiave.
- documento nel quale una determinata data cada in un intervallo temporale assegnato; questa condizione permette di ricercare, ad esempio, i documenti registrati nel flusso diretto o pubblicati in un dato periodo.
- documento nel quale sia presente una determinata parte; a seconda del significato della presenza della parte richiesta, questa "domanda" permette, ad esempio, di individuare i documenti che abbiano superato una certa fase del loro iter o semplicemente quelli che presentino particolari caratteristiche di completezza.

Come risulta chiaro dagli esempi esposti, la funzione della ricerca secondaria è quella di reperire documenti in base a caratteristiche che non si possono descrivere con un insieme di parole-chiave. In conseguenza di ciò, la ricerca secondaria rappresenta, rispetto alla ricerca primaria, un'ulteriore estensione delle possibilità di accesso agli archivi, in quanto "moltiplica" le chiavi dei documenti aumentando in tal modo le probabilità che l'utente possa individuare ciò che cerca sulla sola base delle informazioni in suo possesso e senza dover necessariamente ricorrere a "codici", ossia a dati fissi quali autore, data ecc., la cui conoscenza è spesso problematica.

3. L'estrazione delle parole-chiave

Il sistema individua le parole-chiave nei documenti applicando il trattamento a *lessico aperto* oppure il trattamento a *lessico chiuso.*

Lessico aperto. Questo trattamento richiede la definizione e la memorizzazione preliminare di un archivio di *parole vuote.* Tale archivio contiene tutte le parole (ad sepio congiunzioni, preposizioni, verbi ausiliari ecc.) che non debbono essere prese in considerazione come parole-chiave in quanto non atte ad esprimere una caratteristica del documento. Il sistema estrae le parole-chiave leggendo i documenti parola per parola ed escludendo le parole vuote. Esempio:

- Testo: Legge sulla difesa dell'ambiente
- Parole vuote: sulla, dell
- Parole-chiave: Legge, difesa, ambiente

Lessico chiuso. Questo trattamento richiede la definizione e la memorizzazione preliminare del lessico, ossia di tutte e sole le parole che potranno essere prese in considerazione come parole-chiave. Il sistema estrae le parole-chiave leggendo i documenti parola per parola ed escludendo tutte le parole non comprese nel lessico. Esempio:

- Teso: Legge sulla difesa dell'ambiente
- Lessico: difesa, ambiente
- Parole-chiave: difesa, ambiente

Le parole-chiave possono anche essere stabilite direttamente dall'utente, in tutti i casi in cui non si ritenga opportuno affidarne la scelta ad un trattamento automatico. Nell'ipotesi di scelta manuale delle parole-chiave, l'utente, insieme ai testi dei documenti, deve inserire nel flusso diretto un "riassunto di parole-chiave", che viene elaborato del sistema con tecniche analoghe a quelle impiegate per la scelta automatica.

4. Il thesaurus

Si definisce come *lessico* l'insieme delle parole-chiave che compaiono nel flusso inverso. Tra i termini del lessico è possibile definire delle *relazioni*, il cui utilizzo amplia in modo sostanziale le possibilità di ricerca primaria. Il lessico strutturato con le relazioni prende il nome di *thesaurus.*

Estendere la ricerca primaria ad una relazione significa sostituire ad ogni parola-chiave di una domanda l'unione di tutte le parole-chiave ad essa eventualmente legate da quella relazione.

Continuando l'esempio precedentemente sviluppato, supponiamo di definire una relazione:

- R1 che lega in modo simmetrico tra loro le parole-chiave Aritmetica e Matematica

In questo caso, se la ricerca non fosse estesa, si avrebbe come sempre:

- Domanda: Aritmetica
  Risposta: vol. 1
- Domanda: Matematica
  Risposta: vol. 1, vol. 2, vol. 3, vol. 4

Se invece la ricerca, con un opportuno comando, fosse estesa alla relazione R1, si otterrebbe:

- Domanda: Aritmetica
  Risposta: vol. 1, vol. 2, vol. 3, vol. 4
- Matematica
  Risposta: vol. 1, vol. 2, vol. 3, vol. 4

perché la domanda Aritmetica equivale alla domanda Aritmetica o Matematica e viceversa, la domanda Matematica equivale alla domanda Aritmetica o Matematica.

Le relazioni più importanti e di uso più frequente sono la *sinonimia* e *la gerarchia* (v. figg. 3 e 4).

Sinonimia. La relazione di sinonimia si stabilisce tra parole-chiave che possano considerarsi sinonimi, nel senso usuale del termine. La sinonimia gode delle proprietà simmetrica e transitiva. Se A, B e C sono parole-chiave e se indichiamo col segno = il legame di sinonimia, per la proprietà simmetrica se A = B allora anche B = A, mentre per la proprietà transitiva se A = B e B = C allora anche A = C. Il ri-

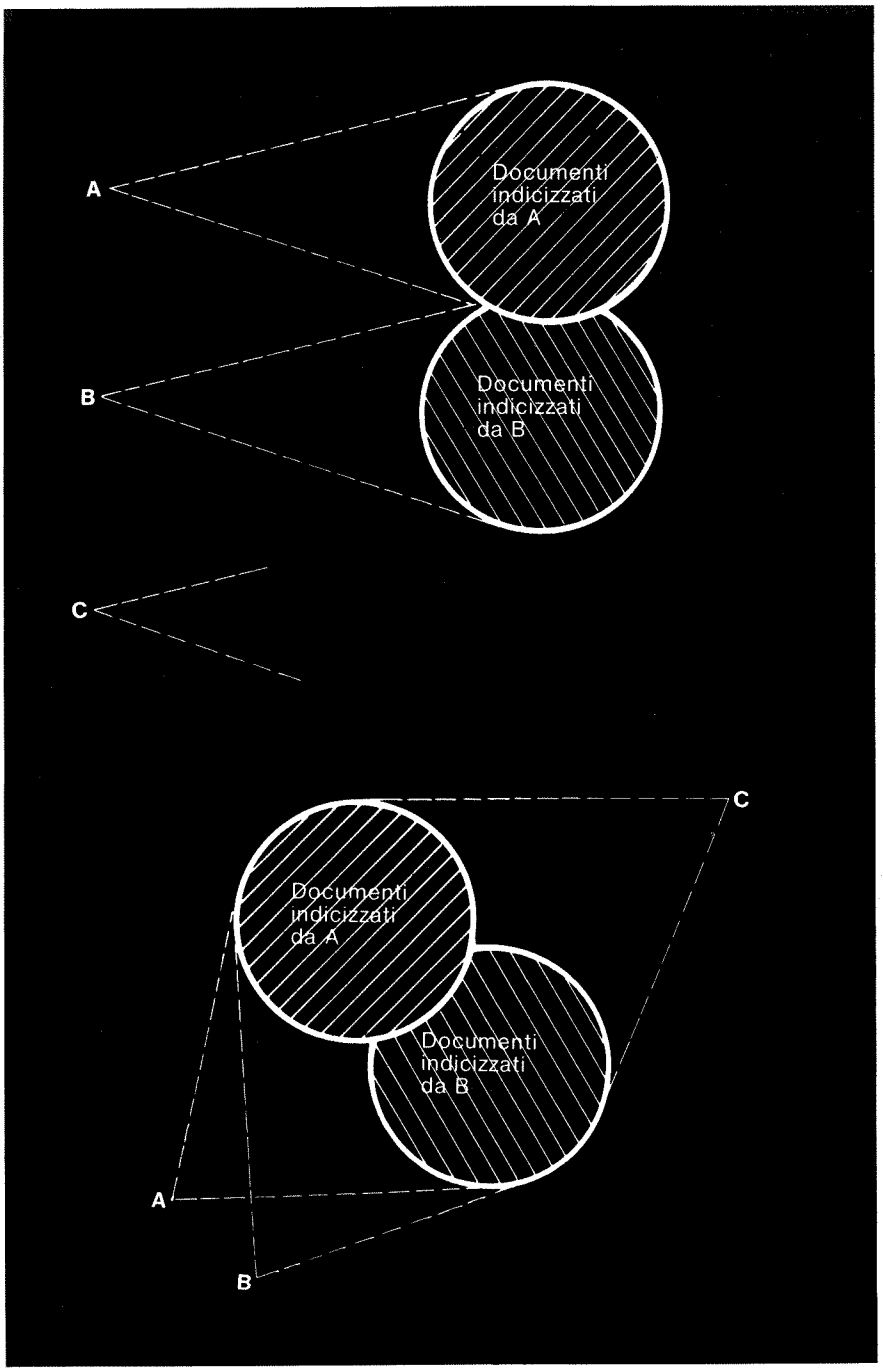


Fig. 3 - Nella figura si illustra graficamente la differenza fra ricerca con lessico senza relazioni e ricerca con thesaurus; si suppone che A, B, C siano tre sinonimi e che C non appartenga al flusso inverso, ossia non indicizzi alcun documento.

spetto di queste proprietà impone che il sistema gestisca un insieme di "richiami" fra le parole-chiave poste in relazione di sinonimia tale che, ad esempio, in caso di ricerca estesa, la domanda A venga trattata come A O B O C, analogamente la domanda B ecc.

Gerarchia. Due parole-chiave sono poste in relazione gerarchica se una costituisce una specificazione dell'altra. Ad esempio le parole *edificio* e *villa* potrebbero essere poste in relazione gerarchica in quanto villa è un particolare tipo di edificio. Delle due parole poste in relazione una si chiama *generico* ed una *specifico*.

Anche la gerarchia prevede una proprietà simmetrica, nel senso che dev'essere possibile andare dallo specifico al generico e viceversa per poter estendere la ricerca alla gerarchia superiore e alla gerarchia inferiore. Analogamente esiste una proprietà transitiva per cui, ad esempio, se A è generico di B e B è generico di C allora anche A è generico di C. Il sistema deve gestire i legami tra le parole in relazione in modo che l'estensione della ricerca possa sfruttare queste proprietà.

L'uso delle relazioni aumenta le probabilità che l'utente possa individuare ciò che cerca ricorrendo a parole-chiave qualsiasi. Infatti la struttura del thesaurus consente di accedere ad un documento non solo attraverso le parole-chiave che lo referenziano nel flusso inverso, ma anche attraverso altre parole-chiave, a quelle legate da qualche relazione. In altri termini, il thesaurus "moltiplica" le parole-chiave che indicizzano i documenti.

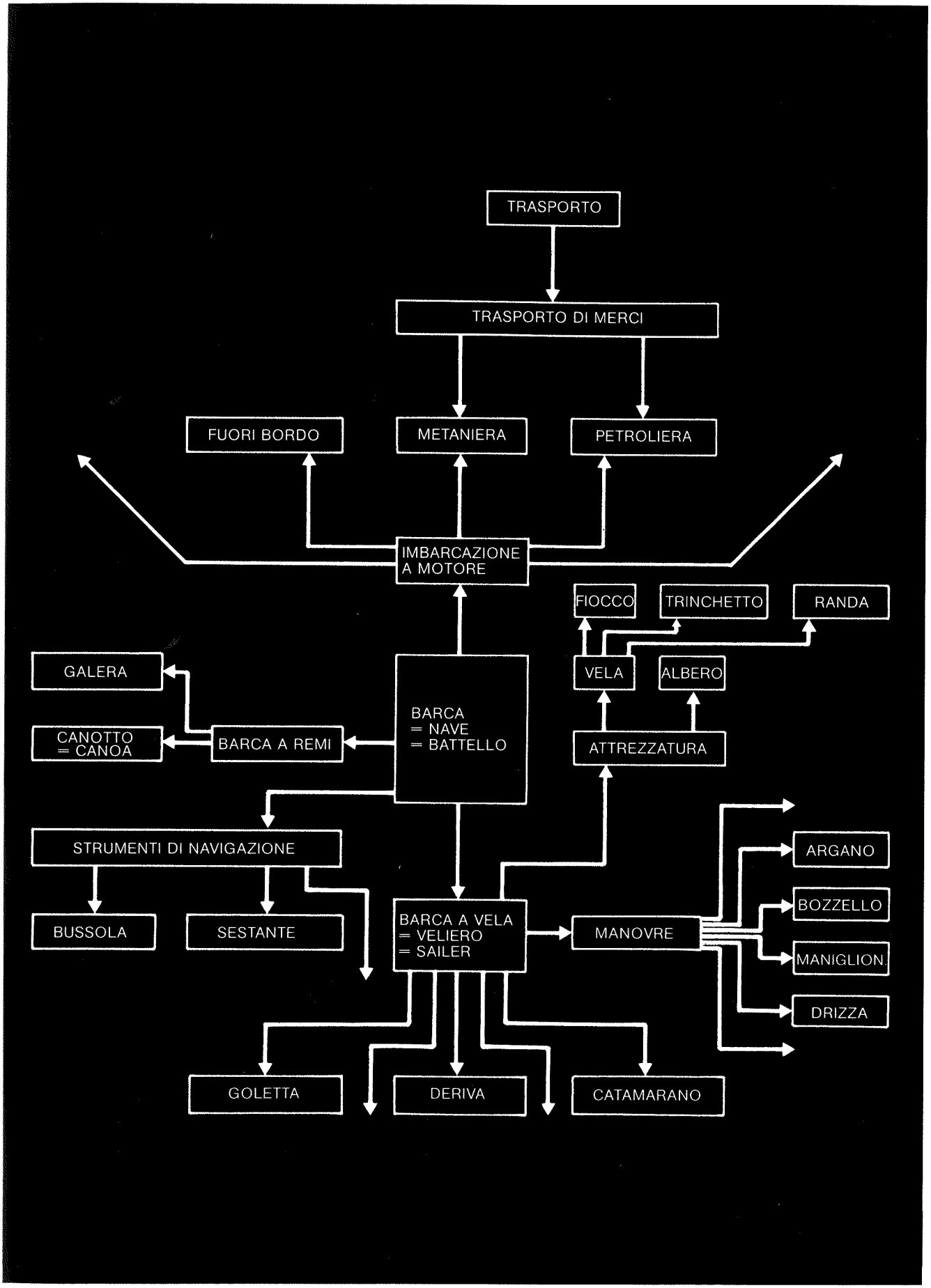


Fig. 4 - Esempio di rappresentazione grafica di un thesaurus con relazioni gerarchiche.

È naturalmente evidente che la struttura del thesaurus, in generale, dev'essere definita "manualmente" dall'utilizzatore, potendo coinvolgere problemi semantici o linguistici che ancora non possono essere affrontati con facilità utilizzando tecniche di elaborazione elettronica dei dati.

5. Considerazioni sui limiti di efficienza dei sistemi di information retrieval

Sintetizzando quanto precedentemente esposto, le possibilità di accesso a un documento del flusso diretto sono:

- le parole-chiave estratte dal documento e/o stabilite dall'utente
- le parole-chiave che non indicizzano quel documento ma sono legate da qualche relazione alle parole-chiave che lo indicizzano
- condizioni varie che caratterizzano il testo del documento e possono essere ricercate con la ricerca secondaria.

Da questa sintesi emerge chiaramente come l'efficienza di un sistema di information retrieval sia condizionata da:

- modalità di determinazione delle parole-chiave
- struttura del thesaurus
- capacità dell'utente di formulare in modo appropriato le proprie domande, sia usando la ricerca primaria che quella secondaria.

È quindi evidente che l'efficienza di un sistema di information retrieval non costituisce solo un problema tecnico, ma dipende anche da altri fattori quali la preparazione dei documenti prima del loro inserimento nel flusso diretto o la dimestichezza degli utenti col "linguaggio delle parole-chiave".

6. Archivio storico Ansaldo: un esempio di applicazione della ricerca documentaria

I sistemi di information retrieval possono essere considerati applicazioni generali, in quanto il loro campo di utilizzazione non è legato a particolari settori di attività nè limitato a tipi specifici di documenti. La ricerca documentaria realizzata con mezzi di elaborazione elettronica è diffusa in:

- biblioteche
- archivi di leggi e/o di sentenze (informatica giuridica)
- reti nazionali o internazionali di diffusione dell'informazione medica, farmaceutica, chimica ecc.
- pubblica amministrazione, per gestire l'accesso a grandi archivi di documenti (p.e. archivio del patrimonio artistico)
- banche
- industria

L'information retrieval rappresenta inoltre una delle funzionalità fondamentali dell'office automation.

Le applicazioni cui accenneremo nel seguito costituiscono di conseguenza semplici esempi tra i tanti che sarebbe possibile produrre.

L'Ansaldo di Genova, fondata nel 1853, è una delle più antiche ed importanti aziende industriali italiane e rappresenta, si può affermare, un luogo privilegiato della nostra storia economica, politica e sociale. Tutte le fasi evolutive del capitalismo italiano, dei suoi rapporti con lo stato e col mondo politico e delle tensioni sociali che ne hanno accompagnato lo sviluppo, sono fedelmente riflesse nelle vicende dell'Ansaldo, dalle quali talvolta hanno avuto inizio processi destinati ad incidere profondamente su tutta la realtà del paese (a questo proposito basti ricordare che l'Ansaldo è stata la prima azienda a partecipazione statale). Una delle eredità di un passato così denso di avvenimenti significativi è costituita da un'immensa mole di documentazione, che dev'essere considerata documentazione storica generale e non semplice documentazione di una storia particolare.

Consapevole dell'importanza di questo materiale e della necessità di renderlo disponibile agli storici, ai ricercatori, agli studenti e a chiunque sia interessato a consultarlo, l'Ansaldo, nel 1977, ha istituito a Genova un Archivio Storico aperto al pubblico, dando così un prezioso contributo allo sviluppo degli studi di storia industriale, in un paese nel quale questa disciplina si trova ancora in uno stato di notevole arretratezza.

Il materiale dell'Archivio Storico Ansaldo è diviso in tre sezioni: fondi archivistici, disegni, fotografie.

La sezione fondi archivistici si compone di un fondo aziendale (circa 300.000 documenti) e di fondi donati da privati. Questi ultimi sono: fondo Mario e Pio Perrone, di circa 700.000 documenti, fondo De Vito, fondo Gamba, fondo Montan, fondo Puri, fondo Rocca, fondo Sarli-Biandrà, fondo Vecoli, fondo Virgilio.

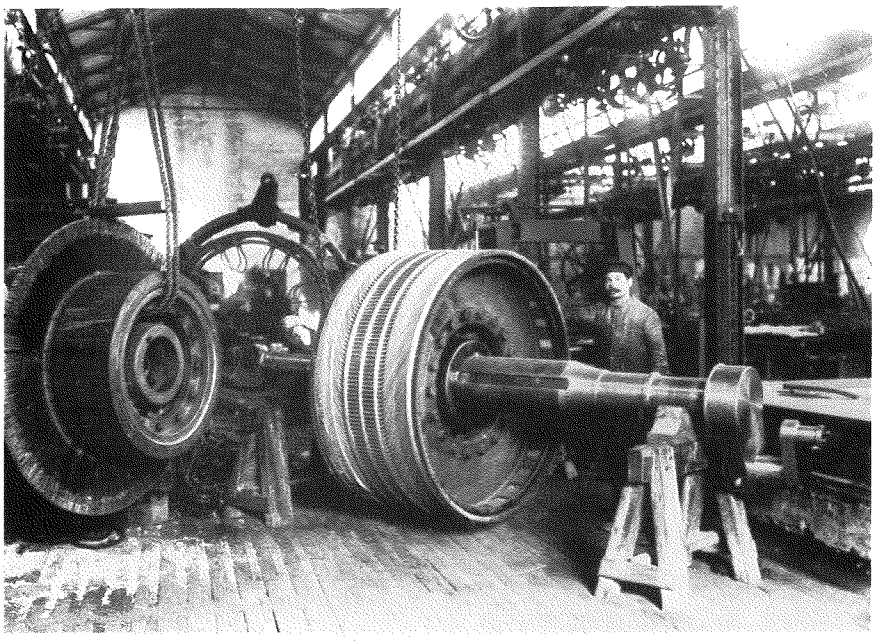
La sezione disegni si compone di oltre 30.000 disegni tecnici di interesse storico relativi a produzione navale, produzione ferroviaria, produzione bellica, produzione meccanica, produzioni minori.

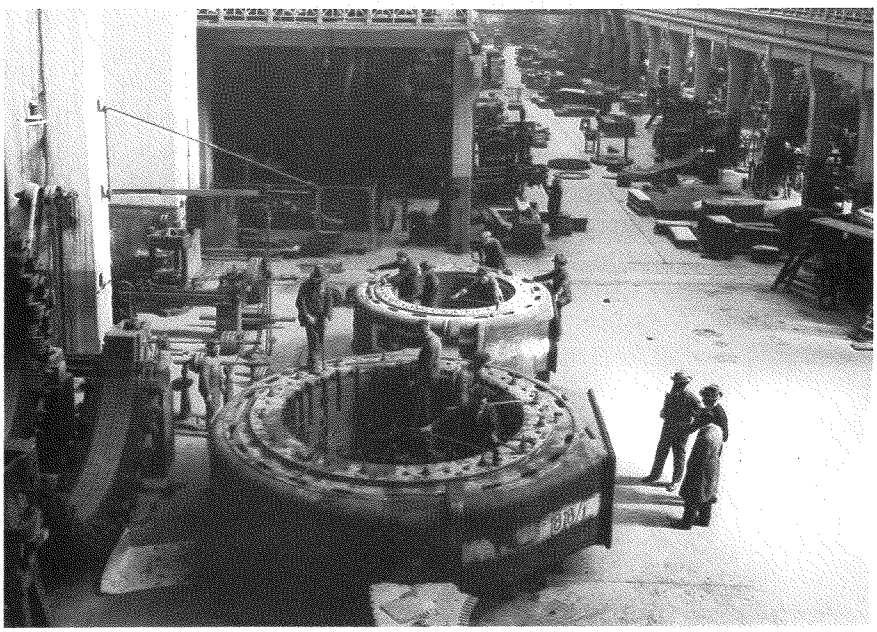
La sezione materiale fotografico riunisce attualmente circa 25.000 negativi, relativi alle molteplici attività dell'Ansaldo.

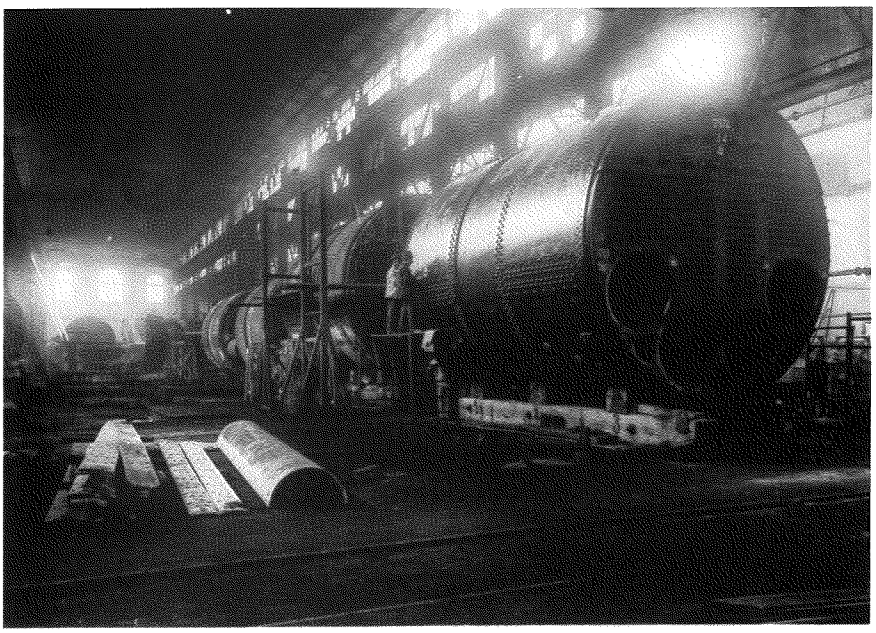
Nel caso dell'Archivio Storico Ansaldo, data la mole di documenti e materiali, si è posto il problema accennato nella parte introduttiva di questo articolo: quali mezzi sono necessari per rendere accessibili le informazioni contenute nell'archivio? Come può l'utente "medio", ossia non esperto dell'archivio, identificare i documenti oggetto della sua ricerca? Il problema è stato affrontato con l'adozione di un sistema di information retrieval della Honeywell, il MISTRAL IV.

In corrispondenza di ogni documento (o insieme di documenti relativi allo stesso oggetto raggruppati in una "cartella"), come di ciascun disegno o fotografia, si registra nel flusso diretto un estratto composto da alcuni dati identificativi essenziali più un insieme di parole-chiave scelte dagli archivisti. Il sistema provvede automaticamente alla gestione del flusso inverso e rende disponibile all'utente finale un indice di notevole efficacia ai fini dell'accessibilità delle informazioni.

Fig. 5 - Alcuni documenti dell'Archivio Storico Ansaldo, nel quale la ricerca è realizzata mediante elaboratore elettronico. Dall'alto in basso:
- lavorazione motori elettrici (1920)
- officina impaccatura lamierini (1926)
- reparto caldereria (1928).







L'applicazione, appartenente al dominio della ricerca bibliografica, pur essendo semplice (infatti non utilizza ad esempio un thesaurus) sta dando ottimi risultati. Tale esito positivo costituisce una valida dimostrazione dell'efficienza del sistema, soprattutto se si considera che l'utenza è veramente "qualsiasi", essendo composta di studiosi, giornalisti, studenti, e comunque di persone che non hanno una specifica conoscenza dell'archivio.

7. L'information retrieval al Ministero della Cultura francese

Il sistema di information retrieval installato presso il Ministero della Cultura francese offre un altro interessante esempio di utilizzazione di queste soluzioni applicative.

Il Servizio Informatico del Ministero della Cultura francese è nato dalla convergenza di tre iniziative:

- Creazione dell'Inventario Generale durante il ministero di André Malraux nel 1964: "censire, studiare, far conoscere" il patrimonio culturale francese in un'ottica di normalizzazione della descrizione e di gestione informatica.
- Creazione di un nuovo museo: il Museo delle Arti e tradizioni Popolari, con la preoccupazione di normalizzare la documentazione sulle collezioni.
- Inizio nel 1971 di uno studio per una gestione più razionale dei crediti riservati ai Monumenti Storici.

Uno dei principali obiettivi di questo servizio è dunque la costituzione delle "banche di dati" relative al patrimonio culturale della Francia:

- opere conservate nelle collezioni pubbliche (musei)
- opere censite nell'Inventario Generale (immobili e oggetti)
- opere protette dal titolo di Monumenti Storici (monumenti e oggetti)
- luoghi archeologici censiti nella "Carta Archeologica"
- fondi particolari degli Archivi di Francia.

Di qui l'importanza dell'*informatica documentaria* (banche di dati e catalogazione) e la scelta di un software capace di gestire queste banche di dati.

La scelta ha condotto all'installazione del prodotto MISTRAL IV nel 1975.

Sono stati messi a punto dei *sistemi descrittivi* in funzione delle possibilità offerte da MISTRAL IV, e sono state create delle banche di dati (pittura, scultura, architettura, oggetti d'arte, antichità egiziane, antichità greco-romane ecc.).

Questi *sistemi descrittivi* sono assai "ricchi", ossia comportano, per ciascuna opera catalogata nel sistema, la registrazione di un considerevole numero di informazioni. Tale "ricchezza" viene completamente sfruttata ai fini della ricerca dalle funzionalità di MISTRAL IV, che consente la ricerca primaria, la ricerca secondaria e gestisce le gerarchie.

Attualmente sono collegati all'elaboratore centrale circa una ventina di terminali.

Diffusione dei piccoli sistemi gestionali nell'industria manifatturiera

DANIELA ROVINA
Honeywell Information Systems Italia
Servizio Ricerche e Sviluppo Marketing
Milano

1. Introduzione

L'utilizzo sempre più capillare di strumenti informatici rappresenta e continuerà a rappresentare negli anni 80 la componente più significativa del settore EDP; l'evoluzione tecnologica avvenuta nell'industria componentistica ha reso possibile la disponibilità di macchine a basso costo e con elevate capacità di elaborazione e ha quindi rivoluzionato il mercato dei piccoli sistemi. Dal lato dell'offerta tale mercato si caratterizza oggi per un elevatissimo numero di fornitori presenti (sino a qualche anno fa nell'ordine della decina) e per il profondo mutamento avvenuto nell'area della distribuzione dei prodotti, di cui il mercato statunitense rappresenta il modello principale. La tipologia degli operatori e dei canali di distribuzione operanti nel settore è, con qualche approssimazione, così sintetizzabile:

- Costruttori di hardware e di tecnologia che rivendono i loro prodotti attraverso sia canali tradizionali (filiali e uffici commerciali) sia rivenditori di apparecchiature e negozi.
- Importatori (di fornitori stranieri) che operano con modeste strutture interne, prevalentemente utilizzate per attività di supporto, e che raramente rivendono i prodotti in forma diretta bensì tramite rivenditori, software house, negozi, ecc.
- Rivenditori di apparecchiature la cui attività principale è costituita dalla vendita dei prodotti, con soluzioni tipo "chiavi in mano", attraverso strutture interne; in taluni casi si avvalgono, a loro volta, per la vendita, di software house, consulenti, negozi, ecc.
- Software house che sviluppano software applicativo più o meno standardizzato per specifici problemi di un determinato settore e che spesso, come già evidenziato, vendono anche hardware, fornendo in tal modo all'utenza soluzioni complete.
- Studi tecnici, consulenti, professionisti vari, che in parallelo alla loro attività specifica vendono anche elaboratori a determinati segmenti di utenza.
- Negozi di prodotti per ufficio, materiale eletronico, HI-FI, ecc., che allo scopo di diversificare la loro attività di vendita al dettaglio, commercializzano anche elaboratori (in prevalenza micro-personal).

È facilmente intuibile che una tale frammentazione dell'offerta comporta una elevata capillarità, sul territorio nazionale, di strutture in grado di offrire e diffondere l'uso di strumenti informatici e creare pertanto i presupposti fondamentali per una reale "informatica di massa".

Per quello che concerne la domanda EDP, da un lato la disponibilità di elaboratori e prezzi a volte inferiori a quelli di un'autovettura, e dall'altro, in una certa misura non meno importante, la facilità d'uso che caratterizza questi sistemi, il fatto, ad esempio, che non richiedono forme di struttura classica EDP (operatori, programmatori, ecc.) e che la gestione del sistema sia delegata a impiegati il cui ruolo professionale è incentrato su mansioni diverse, hanno consentito la diffusione di tecniche informatiche anche in ambienti dove, dati i vincoli strutturali ed organizzativi, l'acquisizione di un elaboratore era considerata un fattore sporadico e casuale.

L'indagine svolta nell'industria manifatturiera (i cui primi e parziali risultati sono stati pubblicati nella "Relazioni e Bilancio 1980 - L'evoluzione del mercato EDP in Italia nel 1980", della Honeywell ISI) si colloca nella necessità di comprendere, in termini più approfonditi, le caratteristiche e le modalità dei piccoli sistemi gestionali nell'ambito di uno dei settori più importanti della economia nazionale. Oggetto della indagine sono le imprese di piccole e medie dimensioni, ovvero con un numero di addetti compreso tra 11 e 250[1]; in tale fascia operano in Italia circa 54 mila aziende che rappresentano, trascurando le im-

prese di tipo artigianale, circa il 96% del settore.

(1) Il campione, di cui era a priori ignoto il grado di meccanizzazione EDP, è stato ottenuto con il metodo di campionamento indiretto. L'errore di stima del campione assicura un C.V. minore o uguale al 10%.
Il riferimento temporale dello studio si colloca nei primi mesi del 1981.

2. Livello e caratteristiche della meccanizzazione

L'uso di strumenti automatici (installati ed in ordine) nell'industria manifatturiera di piccole-medie dimensioni riguarda circa 9400 aziende; il tasso di penetrazione EDP risulta pari al 17.6, mentre il 50.8% delle aziende opera con procedure automatizzate attraverso centro servizi o macchine elettrocontabili.

Il modello di comportamento del fenomeno in esame risulta condizionato principalmente dalle dimensioni aziendali (evidenziate in questa sede come già anticipato, dal numero di addetti); l'uso diretto di un elaboratore riguarda più della metà delle imprese con un numero di dipendenti superiore a 100 (fig. 1) e raggiunge quasi il 35% nella fascia 51-100. L'esigenza della meccanizzazione EDP si delinea comunque, in termini rilevanti, anche in quelle aziende la cui dimensione presenta caratteristiche operative quasi artigianali.

I dati relativi alla meccanizzazione EDP dei singoli sottosettori sembrano riflettere appieno il loro peso economico nell'ambito del totale dell'industria Manifatturiera (fig. 2). I settori Metal-elettromeccanico e Chimico-farmaceutico-gomma manifestano i tassi di penetrazione EDP più elevati (rispettivamente 20.1% e 23.9%); in particolare il Metal-elettromeccanico, data l'elevata numerosità delle imprese in questo settore (in termini percentuali rappresentano, come da fig. 2, il 35% del totale), offre, con circa 15 mila aziende, le più ampie potenzialità di mercato aperto.

L'industria Chimica-farmaceutica-gomma si caratterizza per una elevata dinamicità; dal confronto infatti tra il numero di utenti con elaboratore già installato e quelli con elaboratore in ordine, pur non essendo possibile attribuire una dimensione temporale a tale fenomeno, la crescita che manifesta questo settore è di circa il 30%. Le cause di una dinamica così elevata (il valore medio dell'industria Manifatturiera è del 13.3%) potrebbero essere ricondotte da un lato alla

Classe di addetti	N° totale aziende: struttura percentuale	Tassi di penetrazione EDP
11 - 25	53.1	8.0
26 - 50	26.6	18.7
51 - 100	12.8	34.2
101 - 250	7.5	53.5
Totale	100.0	17.6

Fig. 1 - Tassi di penetrazione EDP per dimensione d'impresa.

Settore merceologico	N° totale aziende: struttura percentuale	Tassi di penetrazione EDP	N° aziende meccanizzate: struttura percentuale
Alimentare	6.9	16.8	6.6
Tessile - abbigliamento	18.7	16.1	17.1
Legno - mobili	10.0	12.6	7.1
Metal - elettromeccanico	35.0	20.1	39.9
Chimico - farmaceutico - gomma	8.0	23.9	10.8
Carta - poligrafiche - editoriali	5.3	18.9	5.7
Resto industria	16.1	14.1	12.9
Totale	100.0	17.6	100.0

Fig. 2 - Tassi di penetrazione EDP per settore merceologico.

Area geografica	N° totale aziende: struttura percentuale	Tassi di penetrazione EDP
Piemonte - Valle d'Aosta - Liguria	12.2	17.7
Lombardia	30.6	21.6
Tre Venezie	15.7	15.6
Emilia-Romagna - Marche	14.1	20.0
Toscana - Umbria	11.0	13.8
Resto Italia	16.4	12.3
Totale	100.0	17.6

Fig. 3 - Tassi di penetrazione EDP per area geografica.

presenza relativamente ampia in questo segmento di aziende di medie dimensioni e dall'altro alla relativa staticità della domanda EDP che il settore ha manifestato negli anni scorsi.

L'elevata concentrazione di aziende di piccolissime dimensioni nel settore Legno-mobili, circa il 60% nella fascia 11-25, unitamente alle particolari caratteristiche strutturali di questo segmento, determinano una scarsa diffusione di tecniche informatiche (il tasso di penetrazione EDP è pari al 12.6%).

Una analisi delle imprese del campione in esame servirà a meglio valutare lo sviluppo dell'uso di strumenti informatici nell'industria Manifatturiera; si è proceduto mantenendo distinto, date le finalità dello studio, il tipo di utenza rilevato in micro, mini e general-purpose[2], anche se, nell'ambito dei piccoli sistemi, tale distinzione risulta spesso di difficile attuazione.

Il 37,8% delle aziende con elaboratore si è meccanizzato, per la prima volta, nel corso del 1980-primi mesi '81; ad una tale dinamica hanno contribuito in modo particolare le aziende la cui fascia dimensionale si colloca fra gli 11 e i 50 addetti: in questo segmento risulta che più della metà degli utenti ha acquisito un elaboratore nel corso del 1980-primi mesi '81 e nell'89% dei casi il tipo di elaboratore scelto era micro o mini. L'immaturità della domanda EDP che questo segmento manifesta viene evidenziata dall'analisi del tipo di utenza per dimensione d'impresa (fig. 4), dove risulta evidente come questa fascia non sia in grado di esprimere un orientamento ben preciso nell'utilizzo di strumenti informatici, in particolare in termini di dimensione del sistema, ma sia largamente condizionata dalle capacità di offerta dei vari fornitori. Prescindendo infatti dall'utenza general-purpose, la cui incidenza percentuale cresce all'aumentare della fascia dimensionale, l'utenza micro e mini manifesta una distribuzione confusa al di sotto dei 50 addetti, ed assume connotazioni più precise solo al di sopra di tale soglia.

Dalla distribuzione percentuale dei sistemi installati (fig. 5) risulta comunque evidente la connessione tra capacità di spesa, esprimibile in termini di dimensione del sistema, e la fascia dimensionale dell'utenza.

Alcune considerazioni circa lo sviluppo settoriale della domanda (figg. 6-7):

- L'industria Tessile-abbigliamento si caratterizza per una elevata diffusione dell'utenza general-purpose (57% contro una media generale del 38%); questa diffusione è tale da consentire al settore, nonostante il suo peso relativo a livello totale, di concentrare il più alto numero di sistemi general-purpose rilevati nell'industria Manifatturiera (29%).
- Prevalenza di utenza micro nei settori Alimentare (58)%, Resto

(2) Senza entrare in una dettagliata descrizione dei criteri adottati, si menziona la corrispondenza fra la classificazione considerata nell'analisi e i prodotti Honeywell:
Micro = Questar M
Mini = Livello 6 - DPS6
General Purpose (piccoli sistemi) = DPS4 - 61 DPS

Fig. 4 - Tipo di meccanizzazione per dimensione d'impresa.

Classe di addetti	Micro	Mini	G.P.	Totale
11 - 25	51	35	14	100
26 - 50	61	19	20	100
51 - 100	29	33	38	100
101 - 250	18	15	67	100
Totale	41	25	34	100

Fig. 5 - Distribuzione percentuale dei sistemi per dimensione d'impresa.

Classe di addetti	Micro	Mini	G.P.	Totale
11 - 25	30	33	10	24
26 - 50	42	21	17	28
51 - 100	18	32	28	25
101 - 250	10	14	45	23
Totale	100	100	100	100

Fig. 6 - Tipo di meccanizzazione per settore merceologico.

Settore merceologico	Micro	Mini	G.P.	Totale
Alimentare	58	5	37	100
Tessile - abbigliamento	27	16	57	100
Legno - mobili	34	13	53	100
Metal - elettromeccanico	47	32	21	100
Chimico - farmaceutico - gomma	33	33	34	100
Carta - poligrafiche - editoriali	24	36	40	100
Resto industria	49	23	28	100
Totale	41	25	34	100

Settore merceologico	Micro	Mini	G.P.	totale
Alimentare	9	1	7	7
Tessile - abbigliamento	11	11	29	17
Legno - mobili	6	4	11	7
Metal - elettromagnetico	46	51	24	40
Chimico - farmaceutico - gomma	9	14	11	11
Carta - poligrafiche - editoriali	3	8	7	5
Resto industria	15	11	11	13
Totale	100	100	100	100

Fig. 7 - Distribuzione percentuale dei sistemi per settore merceologico.

Area	Micro	Mini	G.P.	Totale
Contabilità	97	100	98	98
Fatturazione	85	94	97	91
Gestione ordini clienti	64	88	83	76
Statistiche di vendita	59	75	88	74
Gestione ordini fornitori	51	84	62	62
Paghe e stipendi	66	81	79	74
Magazzino	70	94	95	85
Distinta base	44	63	62	56
Procedure produzione	38	56	53	48

Fig. 8 - Frequenza di automazione delle procedure per tipo di sistema.

Area	Azienda	Fornitore	Altri	Totale
Contabilità	20	57	23	100
Fatturazione	24	51	25	100
Gestione ordini clienti	29	45	26	100
Statistiche di vendita	31	42	27	100
Gestione ordini fornitori	29	46	25	100
Paghe e stipendi	24	53	23	100
Magazzino	26	48	26	100
Distinta base	27	47	26	100
Procedure produzione	33	43	24	100
Totale	26	49	25	100

Fig. 9 - Modalità di sviluppo delle procedure.

Industria (49%) e Metal-elettromeccanico (47%); quest'ultimo, in particolare, concentra il 46% del totale dei sistemi micro e più della metà del totale dei mini rilevati.

Una conferma della maggiore conoscenza, dal lato della domanda, dei vantaggi che l'uso di strumenti automatici apporta alle organizzazioni aziendali, la opportunità della diffusione di tale conoscenza è stata più volte sottolineata dalla letteratura informatica, ci viene fornita dalle imprese del campione, dove risulta che più del 60% ha di sua iniziativa stabilito il primo contatto con il fornitore del sistema. Da un lato si assiste quindi ad una maturazione del mercato il quale, attraverso anche il supporto delle numerose riviste di informatica ormai disponibili e all'attenzione posta all'attività pubblicitaria, particolarmente attiva nell'ultimo periodo, decide in proprio quale può essere lo strumento informatico che meglio si addice alle sue esigenze specifiche; dall'altro, e tale aspetto può considerarsi facente parte della modalità attraverso cui opera l'offerta, il 28% degli utenti è stato sollecitato all'acquisto di un elaboratore solo dopo la visita da parte di un commerciale e il rimanente 11% è venuto in contatto con il fornitore attraverso la partecipazione a dimostrazioni-incontri o l'invio di pubblicazioni promozionali.

I criteri secondo cui l'utenza dell'Industria Manifatturiera decide la scelta di un determinato tipo di elaboratore sono prevalentemente riconducibili alla "serietà e professionalità del fornitore" (35%), sia esso costruttore o rivenditore, ed alla "garanzia di un buon livello di assistenza in caso di guasti" (25%); la disponibilità di pacchetti applicativi offerti dal fornitore non sembra essere un fattore rilevante nella scelta del sistema: solo l'11% l'ha indicato come parametro di scelta principale. Tuttavia sembra importante sottolineare che il 49% degli utenti utilizza unicamente soluzioni applicative preconfezionate ed in particolare, nell'ambito della utenza micro, soltanto il 7% ha sviluppato anche all'interno programmi relativi alle procedure aziendali.

La frequenza di automazione delle principali aree d'attività (figg. 8 e 9) denota livelli applicativi media-

mente piuttosto elevati; ovviamente, date le caratteristiche strutturali delle aziende e le capacità di elaborazione dei sistemi rilevati, le frequenze maggiori di meccanizzazione si riscontrano in corrispondenza delle aree tradizionali di tipo contabile-amministrativo. Alcune considerazioni di carattere più specifico riguardano:

- Il costante sottodimensionamento, rilevato per tutte le aree, dell'utenza micro rispetto a quella mini e general-purpose.
- A livello settoriale, tutti i segmenti manifestano frequenze di automazione delle procedure pressoché costanti; si distingue l'industria Tessile-abbigliamento per un'elevata propensione alla meccanizzazione delle aree relative alla produzione.
- L'utilizzo di pacchetti applicativi sembra riguardare in modo uniforme tutte le aree aziendali, con una maggiore frequenza nelle procedure di tipo contabile-amministrativo.

Un riferimento specifico alle caratteristiche dei sistemi installati nell'industria Manifatturiera concluderà questa panoramica.

Il 74% del totale dei sistemi rilevati nel settore sono stati installati nel biennio '79-'80; si può quindi collocare in questo periodo l'innesco del processo evolutivo che ha condotto il mercato della mini informatica a qualificarsi in termini di "mercato di massa". Nel quadro di tale tendenza complessiva i micro elaboratori denotano una ancor più marcata concentrazione di diffusione nel corso del biennio, l'86% risulta infatti installato nel 1979-1980, e ciò è ovviamente spiegabile se si tiene presente la più recente introduzione sul mercato di tale tipo di prodotto.

La distribuzione delle aziende per numero di posti di lavoro (fig. 10) di cui sono dotati gli elaboratori che esse utilizzano, mette in luce la tendenza ad un utilizzo "conservativo" dello strumento EDP; l'elevata percentuale infatti di calcolatori monostation o comunque con un numero di stazioni di lavoro non superiore a 3 (81%), implica, tendenzialmente, un modello di organizzazione del lavoro in cui l'accesso ai servizi forniti dal calcolatore è ristretto a specifiche persone e/o funzioni aziendali. Vero è, soprattutto per le aziende con meno di 50 addetti, che spesso tutte le attività amministrative in senso lato vengono concentrate da pochissime persone e sono spesso allocate fisicamente in spazi unitari ed è chiaro quindi che in tale contesto organizzativo non si manifestano esigenze di decentramento delle attività di input/output; è anche però vero che non in tutti i casi la spiegazione del fenomeno è riconducibile a tale semplicità organizzativa e denota viceversa un atteggiamento di "prudenza" nei confronti di nuovi strumenti operativi che vengono sì utilizzati ma seguendo approcci e filosofie organizzative di tipo conservativo che guardano con una sosta di diffidenza alla diffusione dello strumento informatico presso tutte le funzioni aziendali.

Problemi di riservatezza, concezioni eccessivamente gerarchiche della struttura d'ufficio oltre, naturalmente, a considerazioni di carattere economico ed ad obbiettive limitazioni delle capacità tecniche (vedi fig. 10) della macchina installata, sconsigliano, almeno in una prima fase, un approccio innovativo che modifichi eccessivamente il tradizionale modus operandi degli uffici amministrativi.

Gli atteggiamenti di prudenza, ovviamente connessi allo sforzo di adattamento richiesto dalla introduzione di nuovi strumenti di lavoro, cederà il passo, man mano che verrà accumulata esperienza in campo EDP, ad una maggiore consapevolezza dei benefici che sono conseguibili della diffusione di strumenti informatici all'interno dell'azienda; oltre quindi che da un'ulteriore diffusione orizzontale di strumenti informatici presso le aziende attualmente sprovviste, le ampissime potenzialità di mercato aperto offerte dall'industria Manifatturiera sono già state evidenziate, il settore trarrà beneficio e occasione di ulteriore rapida espansione dalle accresciute esigenze future degli attuali utilizzatori, che con ogni probabilità, nel giro di qualche anno, tenderanno a migrare verso sistemi informativi più sofisticati e quindi verso macchine con prestazioni maggiori.

Posti di lavoro		Capacità memoria unità centrale		Capacità memoria di massa		Linguaggi utilizzati	
N°	%	KB	%	MB	%	Tipo	%
1	54	≦ 32	41	≦ 2	29	Basic	35
2 - 3	27	32 – 64	26	2 – 10	23	Cobol	11
4 - 5	10	64 – 128	27	> 10	48	RPG	42
≧ 6	9	> 128	6			Altri	12
Totale	100	Totale	100	Totale	100	Totale	100

Fig. 10 - Caratteristiche tecniche dei sistemi.